

LA BIO STATISTIQUE

www.nursunity.ml

Ayoub E. | Mohamed K.

2017



Biostatistique

Sommaire

Biostatistique

Définition :

Biostatistique est un mot-valise issu des champs de la biologie et des statistiques

La Biostatistique est un champ scientifique constitué par l'application de la science statistique à la biologie et à la médecine. ...

Le domaine d'application des biostatistiques est large. Il peut s'agir de biométrie, de conception méthodologique d'études biologiques ou cliniques, ou encore du recueil, de l'analyse et du traitement statistique de données recueillis lors d'études écologiques, biologiques, agronomiques, halieutiques, de santé publique, de santé environnementale d'étudesépidémiologiques, médicales et/ou cliniques, pharmaceutiques, agropharmaceutiques...

Applications

- **Santé publique**

La **santé publique** peut être définie de diverses manières. On peut en effet la présenter comme « l'étude, d'une part, des déterminants physiques, psychosociaux et socioculturels de la santé de la population et d'autre part des actions en vue d'améliorer la santé de la population. »[1] Ou encore, comme « une activité organisée de la société visant à promouvoir, à protéger, à améliorer et, le cas échéant, à rétablir la santé de personnes, de groupes ou de la population entière. »[2]

.La santé publique se démarque de la médecine essentiellement sur deux plans :

1. Elle met davantage l'accent sur la prévention que sur les traitements curatifs ;

2. Elle développe une approche de population, plutôt que de s'intéresser individuellement aux problèmes de santé des personnes, ce qui se traduit notamment par l'adjonction de compétences qui relèvent des sciences humaines et sociales et notamment de la sociologie, du droit et de l'économie.

- **Épidémiologie**

L'**épidémiologie** est l'étude des facteurs influant sur la santé et les maladies de populations. Il s'agit d'une discipline qui se rapporte à la répartition, à la fréquence et à la gravité des états pathologiques.

L'étude de la répartition et des déterminants des évènements de santé sert de fondement à la logique des interventions faites dans l'intérêt de la santé publique et de la médecine préventive.
La reconnaissance de l'épidémiologie comme champ d'étude est relativement récente, puisque la première étude significative remonte à 1854[1] mais elle est l'un des piliers de la santé publique et de la médecine à travers l'histoire.

Les approches des épidémiologistes sont variées : elles vont de l'« étude de terrain » (au cœur de la communauté, souvent dans un service de santé publique) au front de la recherche et de la lutte contre l'émergence des maladies en passant par la modélisation et la veille sanitaire.

➢ Essai clinique

Un **essai clinique**, ou étude clinique, ou encore essai thérapeutique, est une étude scientifique réalisée en thérapeutique médicale humaine pour évaluer l'efficacité et la tolérance d'une méthode diagnostique ou d'un traitement. L'objectif d'un essai n'est pas d'apporter un bénéfice thérapeutique au volontaire[1]. Le Comité international des rédacteurs de revue médicales[2] en donne la définition suivante : « Tout projet de recherche qui affecte de façon prospective des sujets humains à des groupes d'intervention et de comparaison afin d'étudier la relation de cause à effet entre un acte médical et l'évolution d'un état de santé ».

➢ Recherche médicale

La recherche médicale se divise en recherche fondamentale et clinique.

La recherche médicale fondamentale vise à mieux comprendre le corps humain et ses maladies.

La recherche médicale clinique se base sur les résultats de la recherche fondamentale pour inventer et prouver l'efficacité de nouveaux traitements.

➢ Génomique

La génomique est une discipline de la biologie moderne. Elle étudie le fonctionnement d'un organisme, d'un organe, d'un cancer, etc. à l'échelle du génome, au lieu de se limiter à l'échelle d'un seul gène. La génomique se divise en deux branches :

- La *génomique structurale*, qui se charge du séquençage du génome entier ;
- La génomique fonctionnelle, qui vise à déterminer la fonction et l'expression des gènes séquencés en caractérisant le transcriptome et le protéome.

➢ Génétique des populations

La génétique des populations a des applications en épidémiologie où elle permet de comprendre la transmission des maladies génétiques, mais aussi en agronomie, où des programmes de sélection modifient le patrimoine génétique de certains organismes pour créer des races ou variétés plus performantes, ou plus résistantes à des maladies. Elle permet également de comprendre les mécanismes de conservation et de disparition des populations et des espèces (Génétique de la conservation). C'est une discipline des sciences de la vie faisant un fort usage d'outils mathématiques

➢ La statistique

La statistique constitue, en médecine, l'outil permettant de répondre à de nombreuses questions qui se posent en permanence au médecin :

- ✓ Quelle est la valeur normale d'une grandeur biologique, taille, poids, glycémie ?
- ✓ Quelle est la fiabilité d'un examen complémentaire ?
- ✓ Quel est le risque de complication d'un état pathologique, et quel est le risque d'un traitement ?
- ✓ Le traitement A est-il plus efficace que le traitement B ?

1 La variabilité et l'incertain

Toutes ces questions, proprement médicales, reflètent une propriété fondamentale des systèmes biologiques qui est leur variabilité. Cette variabilité est la somme d'une variabilité expérimentale (liée au protocole de mesure) et d'une variabilité proprement biologique. On peut ainsi décomposer la variabilité d'une grandeur mesurée en deux grandes composantes :

$$\text{variabilité totale} = \text{variabilité biologique} + \text{variabilité métrologique}$$

- La variabilité biologique peut être elle-même décomposée en deux termes : d'une part la variabilité intra-individuelle, qui fait que la même grandeur mesurée chez un sujet donné peut être soumise à des variations aléatoires ; et d'autre part la variabilité interindividuelle qui fait que cette même grandeur varie d'un individu à l'autre.

$$\text{variabilité biologique} = \text{variabilité intra-individuelle} + \text{variabilité inter-individuelle}$$

- La variabilité intra-individuelle peut être observée lors de la mesure de la performance d'un athlète qui n'est pas capable des mêmes performances à chaque essai, mais qui se différencie des autres athlètes (variabilité interindividuelle). En général, la variabilité intra est moindre que la variabilité inter.
- La variabilité métrologique peut être elle aussi décomposée en deux termes : d'une part les conditions expérimentales dont les variations entraînent un facteur d'aléas ; et d'autre part les erreurs induites par l'appareil de mesure utilisé.

$$\text{variabilité métrologique} = \text{variabilité expérimentale} + \text{variabilité appareil de mesure}$$

- La mesure de la pression artérielle peut grandement varier sur un individu donné suivant les conditions de cette mesure ; il est ainsi recommandé de la mesurer après un repos d'au moins 15 minutes, allongé, en mettant le patient dans des conditions de calme maximal. Cette recommandation vise à minimiser la variabilité due aux conditions expérimentales. La précision de l'appareil de mesure est une donnée intrinsèque de l'appareil, et est fournie par le constructeur.

2 La mesure d'une grandeur

2.1 Unités et équations aux dimensions

Les grandeurs mesurées possèdent la plupart du temps une unité. La distance entre Paris et Marseille se mesurera par exemple en kilomètres, l'épaisseur d'un cheveu en microns, le poids d'une orange en grammes. Dans quelle mesure ces grandeurs peuvent-elles être comparées ? La distance entre Paris et Marseille, et l'épaisseur d'un cheveu sont deux

longueurs ; leur comparaison est possible si on les mesure avec une unité commune, par exemple le mètre. En revanche, le poids de l'orange n'est pas comparable aux longueurs précédentes.

Deux grandeurs qui peuvent être comparées sont dites posséder la même dimension. Elles peuvent être caractérisées par leur dimension : on parlera par exemple de longueur. Les dimensions de toutes les grandeurs physiques peuvent s'exprimer en fonction de sept dimensions de base : la **longueur** notée **L**, la **masse M**, le **temps T**, l'**intensité électrique I**, la **température Θ**, l'**intensité lumineuse J**, et la **quantité de matière N**.

Par exemple une vitesse est une longueur divisée par un temps. On dira que sa dimension est LT^{-1}.

Plus précisément, de l'équation donnant la vitesse *v* en fonction de la distance *d* parcourue pendant le temps *t*, $v=d/t$, on déduit la relation entre les dimensions (notées entre crochets) des deux membres de l'équation $[v] = [d]/[t] = LT^{-1}$. Cette relation est appelée équation aux dimensions.

Une équation aux dimensions permet donc d'exprimer la dimension de n'importe quelle grandeur en fonction des dimensions élémentaires, à condition de connaître les relations entre elles. Elle permet aussi une première validation d'une relation entre grandeurs physiques : les dimensions de la partie gauche et de la partie droite de la relation doivent être identiques.

Déterminons par exemple la dimension d'une énergie ou d'un travail, à partir de la formule $w = f.l$ (un travail est le produit d'une force par une longueur). Une force est le produit d'une masse par une accélération ($f = m.\gamma$) et une accélération est une longueur divisée par le carré d'un temps. Donc $[w] = [f][l] = [m][l]t^{-2}][l] = ML^2T^{-2}$.

Un autre intérêt des équations aux dimensions concerne les unités des grandeurs mesurées. On définit un système d'unités en imposant des unités aux 7 dimensions de base, les autres unités de définissant à l'aide des équations aux dimensions. Le système d'unités le plus utilisé est le **Système International**, ou SI, dans lequel une longueur est mesurée en **mètres** (m), une masse en **kilogrammes** (kg), un temps en **secondes** (s), une intensité électrique en**ampères** (A), une température en **degrés Kelvin** (K), une intensité lumineuse en **candelas** (cd), et une quantité de matière en **moles** (mol).

Dans le système international, certaines unités dérivées sont évidentes : une surface s'exprime en mètres carrés. D'autres le sont moins. Citons l'*hertz* pour une fréquence, le *pascal* pour une pression, le *joule* pour une énergie ou un travail, le *watt* pour une puissance, le *newton* pour une force, le *coulomb* pour une charge électrique, le *volt* pour une différence de potentiel, l'*ohm* pour une résistance, etc.

Il existe des grandeurs sans dimension, calculées comme le rapport de deux grandeurs de même dimension, mais qui possèdent pourtant une unité. Un angle est une grandeur sans dimension mesurée en *radians* dans le système international.

2.2 Erreurs de mesure

La mesure d'une grandeur ne peut conduire à une valeur exacte. En premier lieu, l'instrument de mesure possède nécessairement une précision limitée : une règle graduée millimètre par millimètre ne peut donner une meilleure précision qu'un demi millimètre. En second lieu, la grandeur à mesurer peut être source de variabilité intra-individuelle : la répétition de la mesure avec le même instrument et dans des conditions identiques conduit alors à des résultats différents. Enfin, l'instrument de mesure peut être mal étalonné ou mal adapté et conduire à un biais de mesure systématique : les valeurs mesurées seront systématiquement trop élevées, ou systématiquement trop basses.

Pour une grandeur X à mesurer, on note ΔX l'erreur de mesure. Cette erreur est généralement facilement connue si elle n'est due qu'à un problème de précision. S'il existe une variabilité intra-individuelle (raisonnablement faible), on fera intervenir l'écart-type des mesures (voir chapitre 10 « Estimation - Intervalle de confiance »). Si x est la valeur mesurée, la vraie valeur est donc comprise entre $x-\Delta X$ et $x+\Delta X$.

Si une grandeur G n'est pas mesurée, mais déduite d'autres grandeurs X, Y, Z à l'aide d'une formule, l'erreur ΔG sur Gdoit se déduire des erreurs ΔX, ΔY, ΔZ sur X, Y, Z.

Le plus souvent, on utilise un calcul basé sur la différentielle totale exacte de la formule. Si $G = f(X, Y, Z)$, la différentielle totale exacte[1] est :

$$df = \frac{\partial f}{\partial X}dX + \frac{\partial f}{\partial Y}dY + \frac{\partial f}{\partial Z}dZ$$

L'erreur de mesure est alors donnée par :

$$\Delta G = \Delta f = \left|\frac{\partial f}{\partial X}\right|\Delta X + \left|\frac{\partial f}{\partial Y}\right|\Delta Y + \left|\frac{\partial f}{\partial Z}\right|\Delta Z$$

Supposons par exemple devoir calculer une résistance R en mesurant l'intensité I du courant qui y circule et la différence de potentiel U à ses bornes. La formule liant ces grandeurs est $R = U/I$. On mesure $U = 1000$ volts à 1 volt près et $I = 1$ ampère à 10^{-3} ampère près.

La formule donne $R = 1000$ ohms et l'erreur se calcule par $\Delta R = \frac{\Delta U}{I} + \frac{U}{I^2}\Delta I = 2$

Le calcul basé sur la différentielle totale exacte n'est cependant qu'une approximation (on confond une courbe et sa tangente). Lorsque des calculs plus exacts sont possibles, ils sont préférables. Ainsi, supposons avoir trouvé 100 avec une précision de 1 pour la mesure d'une grandeur X et nous intéresser à la grandeur $Y = 1/X$.

La formule de la différentielle totale exacte donne $\Delta Y = \frac{\Delta X}{X^2} = 10^{-4}$, donc une valeur de Y comprise entre 0,0099 et 0,0101.

Mais puisque la vraie valeur de X est comprise entre 99 et 101, la vraie valeur de Y est en réalité comprise entre 1/101 et 1/99, soit entre 0,009901 et 0,010101.

3 La décision dans l’incertain

Pour prendre une décision diagnostique ou thérapeutique le médecin doit avoir des éléments lui permettant de prendre en compte cette variabilité naturelle, pour distinguer ce qui est normal de ce qui est pathologique (décision à propos d’un patient) et pour évaluer la qualité d’un nouvel examen, ou d’une nouvelle thérapeutique (décision thérapeutique). La compréhension des méthodes statistiques, de leur puissance et de leurs limites, est essentielle pour un médecin de nos jours. Tout résultat de recherche médicale résulte d’une expérimentation (clinique ou biologique) qui s’appuie sur une méthodologie statistique rigoureuse, et dont les résultats sont analysés en termes statistiques.

De même la démarche statistique permet d’évaluer les risques (ou les bénéfices) d’une prescription, de déterminer dans une situation donnée l’examen qui apportera la meilleure information diagnostique.

Nous voyons donc l’importance de la maîtrise de l’outil et de la démarche statistique :

- Pour permettre les progrès de la connaissance médicale : c’est le domaine de la recherche clinique qui ne peut s’accomplir convenablement (définition de la question, mise en place du protocole expérimental, analyse des résultats) qu’en suivant une méthodologie statistique rigoureuse.
- Pour mieux connaître l’état de santé d’une population, la fréquence et la gravité d’une épidémie (penser au SIDA), etc. Cette connaissance se fera à partir d’échantillons convenablement choisis et de calculs basés sur les outils de la statistique. Il sera alors possible de rechercher les stratégies de prévention les mieux adaptées, d’en évaluer leur impact. Il s’agit là des applications relevant de l’épidémiologie et de la santé publique.
- Pour améliorer la pratique médicale dans ses aspects décisionnels, à savoir choisir le meilleur examen (clinique ou para-clinique) pour aboutir le plus rapidement et le plus sûrement au diagnostic. Pour optimiser la thérapeutique, choisir le traitement le mieux adapté à un patient donné (choix du médicament, posologie, etc).

L’objectif de ce cours est de vous fournir les bases indispensables permettant de comprendre les méthodes utilisées, d’interpréter correctement les résultats de nouvelles recherches, et d’adopter un mode de raisonnement qui soit à même d’aider à la décision dans l’exercice de la médecine.

Plus précisément nous étudierons successivement :

1. Les bases de calcul de probabilités, qui sont indispensables à la compréhension et à l’utilisation des méthodes statistiques.
2. La statistique descriptive qui permet de représenter et de quantifier la variabilité d’une ou plusieurs grandeurs observées.
3. La statistique inductive qui inclura les tests statistiques permettant de retenir une hypothèse A plutôt qu’une hypothèse B à partir de données expérimentales (comme dans le cas de la comparaison de deux traitements, où l’hypothèse A est que les deux traitements sont équivalents et l’hypothèse B est qu’ils sont différents).
4. Les applications des méthodes statistiques à l’épidémiologie, à l’aide à la décision thérapeutique et diagnostique, et les applications aux essais thérapeutiques.

Chapitre 1 : Statistique(s) et Probabilité(s)

Nous commencerons par définir les termes et les concepts importants.

1.1 Statistique

Le terme statistique désigne à la fois un ensemble de données d'observations, et l'activité qui consiste en leur recueil, leur traitement et leur interprétation. Les termes *statistique, ou statistiques* (au pluriel) englobent ainsi plusieurs notions distinctes :

1. D'une part le recensement de grandeurs d'intérêt comme le nombre d'habitants d'un pays, le revenu moyen par habitant, le nombre de séropositifs dans la population française. Nous voyons que la notion fondamentale qui se dégage de cette énumération est celle de *Population*. Une population est un ensemble d'objets, d'êtres vivants ou d'objets abstraits (ensemble des mains de 5 cartes distribuées au bridge...) de même nature.
2. La statistique en tant que science s'intéresse aux propriétés des populations naturelles. Plus précisément elle traite de nombres obtenus en comptant ou en mesurant les propriétés d'une population. Cette population d'objets doit en outre être soumise à une variabilité, qui est due à de très nombreux facteurs inconnus (pour les populations d'objets biologiques qui nous intéressent ces facteurs sont les facteurs génétiques et les facteurs environnementaux).
3. A ces deux acceptions du terme *statistiques* (au pluriel) il faut ajouter le terme *statistique* (au singulier) qui définit toute grandeur calculée à partir d'observations. Ce peut être la plus grande valeur de la série statistique d'intérêt, la différence entre la plus grande et la plus petite, la valeur de la moyenne arithmétique de ces valeurs, etc.

1.2 Population et échantillon

On appelle *population* P un ensemble généralement très grand, voire infini, d'individus ou d'objets de même nature. Tous les médecins de France constituent une population, de même que l'ensemble des résultats possibles du tirage du loto. Une population peut donc être réelle ou fictive.

Il est le plus souvent impossible, ou trop coûteux, d'étudier l'ensemble des individus constituant une population ; on travaille alors sur une partie de la population que l'on appelle *échantillon*. Pour qu'un échantillon permette l'étude de la variabilité des caractéristiques d'intérêt de la population, il faut qu'il soit convenablement sélectionné. On parlera*d'échantillon représentatif* si les individus le constituant ont été tirés au sort[1] dans la population. Si par exemple on souhaite déterminer les caractéristiques « moyennes » du poids et de la taille des prématurés masculins on tirera au hasard un certain nombre de sujets parmi les naissances de prématurés de l'année.

Chaque individu, ou unité statistique, appartenant à une population est décrit par un ensemble de caractéristiques appelées *variables* ou *caractères*. Ces variables peuvent être quantitatives (numériques) ou qualitatives (non numériques) :

Quantitatives

Pouvant être classées en variables continues (taille, poids) ou discrètes (nombre d'enfants dans une famille)

Qualitatives

Pouvant être classées en variables catégorielles (couleurs des yeux) ou ordinales (intensité d'une douleur classée en nulle, faible, moyenne, importante).

1.3 Statistique et probabilité

La théorie (ou le calcul) des probabilités est une branche des mathématiques qui permet de modéliser les phénomènes où le hasard intervient (initialement développée à propos des jeux de hasard, puis progressivement étendue à l'ensemble des sciences expérimentales, dont la physique et la biologie).

Cette théorie permet de construire des modèles de ces phénomènes et permet le calcul : c'est à partir d'un modèle probabiliste d'un jeu de hasard comme le jeu de dés que l'on peut prédire les fréquences d'apparition d'événements comme le nombre de fois que l'on obtient une valeur paire en jetant un dé un grand nombre de fois. Les éléments de calcul des probabilités indispensables à la compréhension des statistiques seront traités dans la première partie du cours.

Sous jacente à la notion de statistiques se trouve la notion de Population dont on souhaite connaître les propriétés (plus précisément les régularités), permettant en particulier de savoir si deux populations sont identiques ou non. Ce cas est celui du cadre des essais thérapeutiques, où l'on considère 2 populations (patients traités avec le médicament A ou avec le médicament B) dont on souhaite savoir si elles diffèrent ou non (c'est le cas le plus simple des essais cliniques). Pour ce faire il est nécessaire de modéliser les populations, en utilisant des modèles probabilistes. Un modèle de ce type est par exemple de considérer que la taille des individus suit une distribution gaussienne. A partir de ce modèle on peut calculer les propriétés d'échantillons ; c'est ce qu'on appelle une déduction qui va du modèle vers l'expérience. A l'inverse, considérant un échantillon d'une population on peut essayer de reconstruire le modèle de la population.

Cette démarche est calquée sur la démarche scientifique habituelle. Le scientifique est capable, en utilisant les mathématiques, de prédire le comportement d'un modèle donné (c'est par exemple une « loi » de la physique) : c'est la démarche déductive. A l'inverse, observant des faits expérimentaux il va tenter de dégager des propriétés générales du phénomène observé qu'il va en général représenter sous forme d'un modèle (toutes les lois de la physique et de la chimie sont des modèles mathématiques les plus généraux possibles des faits expérimentaux) : c'est la construction inductive de la théorie. Cette démarche générale va plus loin car le modèle permet de prédire des expériences non réalisées. Si les prédictions ainsi réalisées sont contradictoires avec les résultats expérimentaux alors on pourra avec certitude réfuter le modèle (on dit aussi qu'on l'a falsifié) ; dans le cas contraire on garde le modèle mais on n'est pas certain qu'il soit « vrai ». Autrement dit, à l'issue d'un tel test on ne peut avoir de certitude que si on a trouvé des éléments permettant de réfuter le modèle. Nous verrons dans la suite que cette approche se transpose exactement dans la démarche statistique, en particulier dans le domaine des tests.

Chapitre 2 : Rappels mathématiques

2.1 Ensembles, éléments

On appelle *ensemble*, toute liste ou collection d'objets bien définis, explicitement ou implicitement ; on appelle éléments ou membres de l'ensemble les objets appartenant à l'ensemble et on note :

- $p \in A$ si *p* est un élément de l'ensemble *A*
- *B* est partie de *A*, ou sous ensemble de *A*, et l'on note $B \subset A$ ou $A \supset B$, si $x \in B \Rightarrow x \in A$

On définit un ensemble soit en listant ses éléments, soit en donnant la définition de ses éléments :

- $A = \{1, 2, 3\}$
- $X = \{x : x$ est un entier positif$\}$

Notations :

- la négation de $x \in A$ est $x \notin A$
- Ø est l'ensemble vide
- *E* est l'ensemble universel.

2.2 Opérations sur les ensembles

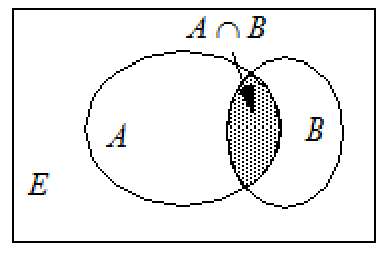
Soient *A* et *B* deux ensembles quelconques.

Intersection

L'intersection de *A* et *B*, notée $A \cap B$, est l'ensemble des éléments *x* tels que $x \in A$ et $x \in B$. Soit :

$A \cap B = \{ x : x \in A$ et $x \in B \}$

Le terme « et » est employé au sens $x \in A$ et B si *x* appartient à la fois à *A* et à *B*



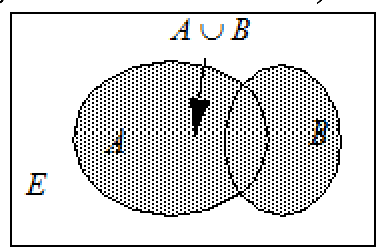
Cas particulier : si $A \cap B = \varnothing$, on dit que *A* et *B* sont **disjoints**.

Réunion

La réunion de *A* et *B*, notée $A \cup B$, est l'ensemble des éléments *x* tels que $x \in A$ ou $x \in B$. Soit :

$A \cup B = \{ x : x \in A$ ou $x \in B \}$

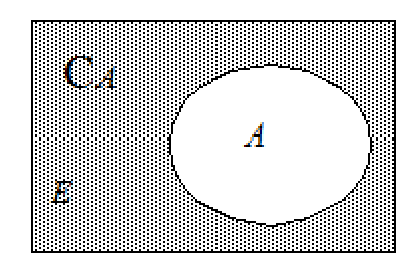
Le terme « ou » est employé au sens $x \in A$ ou B si x appartient à A, ou à B, ou à A et B (car $x \in A$ et B signifie $x \in A$ et $x \in B$).



Complémentaire

Le complémentaire de A est l'ensemble des éléments de E qui n'appartiennent pas à A.

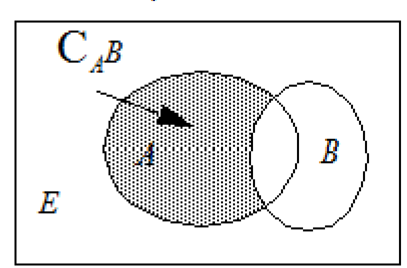
$$\complement_A = \bar{A} = \{x : x \notin A\}$$



Différence

La différence entre A et B, ou complémentaire de B relatif à A, est l'ensemble des éléments de A qui n'appartiennent pas à B.

$$A - B = \complement_A B = \{x : x \notin B \text{ et } x \in A\}$$



Algèbre des ensembles

$A \cup A = A$	$A \cap A = A$
$(A \cup B) \cup C = A \cup (B \cup C)$	$(A \cap B) \cap C = A \cap (B \cap C)$
$A \cup B = B \cup A$	$A \cap B = B \cap A$
$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
$A \cup \varnothing = A$	$A \cap E = A$
$A \cup E = E$	$A \cap \varnothing = \varnothing$
$A \cup \complement A = E$	$A \cap \complement A = \varnothing$
$\complement\complement A = A$	$\complement E = \varnothing$, $\complement \varnothing = E$
$\complement(A \cup B) = \complement A \cap \complement B$	$\complement(A \cap B) = \complement A \cup \complement B$

Chapitre 3 : Eléments de calcul des Probabilités

3.1 Introduction

Le calcul des probabilités est la théorie mathématique, donc fondée axiomatiquement, qui permet de modéliser des phénomènes aléatoires, ou non déterministes.

De tels phénomènes sont bien représentés par les jeux de hasard dont l'étude a initié le calcul des probabilités. Considérons le cas du jeu de dés ; lorsqu'on jette un dé on est certain qu'il va tomber sur la table (phénomène déterministe), mais on n'est pas capable de prédire la valeur qui va sortir (phénomène aléatoire).

Un phénomène déterministe est un phénomène dont on peut prévoir le résultat ; les lois de la physique classique sont des modèles permettant de prédire le résultat d'une expérience donnée. La loi d'Ohm permet de prédire la valeur de l'intensité du courant connaissant la résistance et la tension aux bornes. Les lois de la physique mettent en évidence une régularité qui permet de prédire les résultats d'une expérience lorsqu'on contrôle les causes. Les phénomènes aléatoires exhibent un autre type de régularité. Prenons le cas des lois de Mendel. Mendel était un biologiste qui étudiait les résultats du croisement de deux espèces de plantes ; plus précisément, il étudiait la transmission de caractères comme la couleur, l'aspect, etc. Une observation typique de régularité d'un nouveau type est d'observer que, sur une série suffisamment grande de croisements de deux espèces A et B, on observait par exemple, dans 1/4 des cas, les caractères de A, et dans 3/4 des cas, les caractères de B. Une telle régularité fréquentielle a donné lieu à ce qu'on appelle les lois de Mendel. Cette régularité permet de prédire la fréquence d'apparition d'un phénomène, ce qui est plus « faible » que la prédiction déterministe. L'étude et la modélisation de tels phénomènes (la recherche de lois) est le champ d'application du calcul des probabilités.

3.2 Expérience aléatoire, ensemble fondamental et événements

Expérience aléatoire

On s'intéresse ici aux seules expériences dont le résultat n'est pas prévisible, les expériences aléatoires. Une expérience aléatoire est aussi appelée une *épreuve*.

Ensemble fondamental

Pour une expérience aléatoire donnée, l'ensemble des résultats possibles est appelé l'ensemble fondamental, que nous noterons E dans la suite du cours.

Chaque **résultat** d'expérience est un point de E ou un élément de E.

Evénement

Un événement A est un sous ensemble de E, c'est-à-dire un ensemble de résultats.

L'événement $\{a\}$, constitué par un seul point de E, donc par un seul résultat $a \in E$, est appelé **événement élémentaire**.

L'ensemble vide Ø ne contient aucun des résultats possibles : il est appelé **événement impossible**.

L'ensemble E contient tous les résultats possibles : c'est l'**événement certain**.

Si E est fini, ou infini dénombrable, tout sous-ensemble de E est un événement ; ce n'est pas vrai si E est non dénombrable (ceci sort du cadre de ce cours).

On note parfois Ω l'ensemble de tous les événements.

Exemples

1. On jette un dé et on observe le résultat obtenu. L'ensemble fondamental est formé par les 6 résultats possibles :
   $E = \{1, 2, 3, 4, 5, 6\}$
   L'événement correspondant à l'apparition d'un nombre pair est $A = \{2, 4, 6\}$, qui est

bien un sous ensemble de E.
L'événement correspondant à l'apparition d'un nombre premier est $B = \{1, 2, 3, 5\}$, et l'événement correspondant à l'apparition d'un 3 est $C = \{3\}$.

2. Dans l'exemple précédent E était fini et donc dénombrable ; E peut être infini dénombrable comme dans le cas suivant. On jette une pièce de monnaie jusqu'à ce qu'on obtienne pile ; l'ensemble fondamental correspondant est la suite des nombres entiers $E = \{1, 2, 3, ..., n, ...\}$ puisqu'on peut avoir un pile au bout d'un jet, de 2 jets, de n jets, n étant aussi grand que l'on veut.
3. On vise avec une fléchette une cible suffisamment grande ; si on admet que la fléchette est très fine, comme le serait un point de la géométrie, l'espace fondamental est la surface de la cible qui est constituée de points et donc infinie et non dénombrable.

3.3 Opérations sur les événements

Les événements peuvent se combiner entre eux pour former de nouveaux événements.
Si A et B sont deux événements, les opérations de combinaison sont :

1. $A \cup B$ est l'événement qui se produit si A ou B (ou les deux) est réalisé.
   Il est parfois noté $A + B$ ou A ou B.
2. $A \cap B$ est l'événement qui se produit si A et B sont réalisés tous les deux.
   Il est parfois noté $A \cdot B$ ou A et B.
3. $\complement A$ est l'événement qui se produit quand A n'est pas réalisé. On l'appelle aussi négation de A.
   Il est parfois noté « $\text{non } A$ », ou $\bar{A}$.

Evénements incompatibles

Quand deux événements A et B sont tels que $A \cap B = \varnothing$, ils ne peuvent être réalisés simultanément. On dit qu'ils s'**excluent mutuellement**, ou qu'ils sont **incompatibles**.

Système complet d'événements

On dit que les événements $A_1, A_2, ..., A_n$ forment une famille complète si les A_i constituent une partition de E, c'est-à-dire si :

1. les événements sont deux à deux disjoints : $\forall (i \neq j), (A_i \cap A_j = \varnothing)$
2. ils couvrent tout l'espace : $\bigcup_i A_i = E$

Exemple

Reprenons l'exemple précédent du jeu de dés :
$E = \{1, 2, 3, 4, 5, 6\}$, $A = \{2, 4, 6\}$, $B = \{1, 2, 3, 5\}$, $C = \{3\}$.
$A \cup B = \{1, 2, 3, 4, 5, 6\}$ = apparition d'un nombre pair ou premier
$A \cap B = \{2\}$ = apparition d'un nombre pair et premier
$\complement C = \{1, 2, 4, 5, 6\}$ = apparition d'un nombre autre que 3
$A \cap C = \varnothing$: A et C s'excluent mutuellement.

Chapitre 4 : Probabilité Conditionnelle ; Indépendance et Théorème de Bayes

4.1 - Probabilité conditionnelle

Soient *A* et *B* deux événements quelconques d'un ensemble fondamental *E* muni d'une loi de probabilité *Pr*. On s'intéresse à ce que devient la probabilité de *A* lorsqu'on apprend que *B* est déjà réalisé, c'est-à-dire lorsqu'on restreint l'ensemble des résultats possibles *E* à *B*.

La probabilité conditionnelle de *A*, sachant que l'événement *B* est réalisé, est notée *Pr*(*A*/ *B*) et est définie par la relation suivante :

$$Pr(A/B) = \frac{Pr(A \cap B)}{Pr(B)}$$

Equation 1 : probabilité conditionnelle

Dans cette équation, les probabilités des événements $A \cap B$ et *B* doivent être calculées sur tout l'ensemble fondamental *E*, comme si on ne savait pas que *B* s'est déjà réalisé. Sinon, on obtient évidemment *Pr*(*B*) = 1.

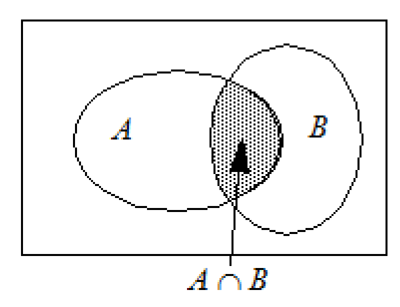


Figure 1 : probabilité conditionnelle

Cette relation générale pour tout espace probabilisé s'interprète facilement dans le cas où *E* est un espace équiprobable (mais cette relation est vraie pour un espace non-équiprobable !). En notant $|A|$ le nombre d'éléments de*A* :

$$Pr(A \cap B) = \frac{|A \cap B|}{|E|},\ Pr(B) = \frac{|B|}{|E|},\ Pr(A/B) = \frac{|A \cap B|}{|B|}$$

Pr(*A*/ *B*) traduit le rapport de la surface de $A \cap B$ sur la surface de B dans la figure 1.

Toujours dans le cas où *E* est équiprobable, on a

$$Pr(A/B) = \frac{\text{nombre de réalisations possibles de } A \text{ et } B \text{ en même temps}}{\text{nombre de réalisations de } B}$$

Cette interprétation de la probabilité conditionnelle, facile à appréhender dans le cas d'équiprobabilité, est la définition générale de la probabilité conditionnelle qu'on doit utiliser telle quelle, sans chercher une interprétation fréquentiste dans tous les cas.

Exemple

On jette une paire de dés bien équilibrés (espace équiprobable). On observe une réalisation de l'événement {somme des dés = 6}. Quelle est la probabilité pour qu'un des deux dés ait donné le résultat 2 ?
B = {somme des deux dés = 6}
A = {au moins un des deux dés donne 2}
B = {(2, 4), (4, 2), (1, 5), (5, 1), (3, 3)}
Nombre de réalisations de $A \cap B$ = {(2, 4), (4, 2)} = 2

D'où $Pr(A/B) = \frac{|A \cap B|}{|B|} = \frac{2}{5}$, alors que $Pr(A) = \frac{11}{36}$

4.2 Théorème de la multiplication

Reprenons l'équation 1, définition des probabilités conditionnelles : $Pr(A/B) = \frac{Pr(A \cap B)}{Pr(B)}$

On en tire immédiatement

$$Pr(A \cap B) = Pr(A/B)Pr(B) = Pr(B/A)Pr(A)$$

Equation 2 : théorème de la multiplication

L'équation 2 peut se généraliser facilement. Soient A_1, ..., A_n des événements quelconques d'un espace probabilisé ; à partir de l'équation 2, on montre :

$$Pr(A_1 \cap A_2 \cap \ldots \cap A_n) = Pr(A_1)Pr(A_2/A_1)Pr(A_3/(A_1 \cap A_2))\ldots Pr(A_n/(A_1 \cap A_2 \ldots \cap A_{n-1}))$$

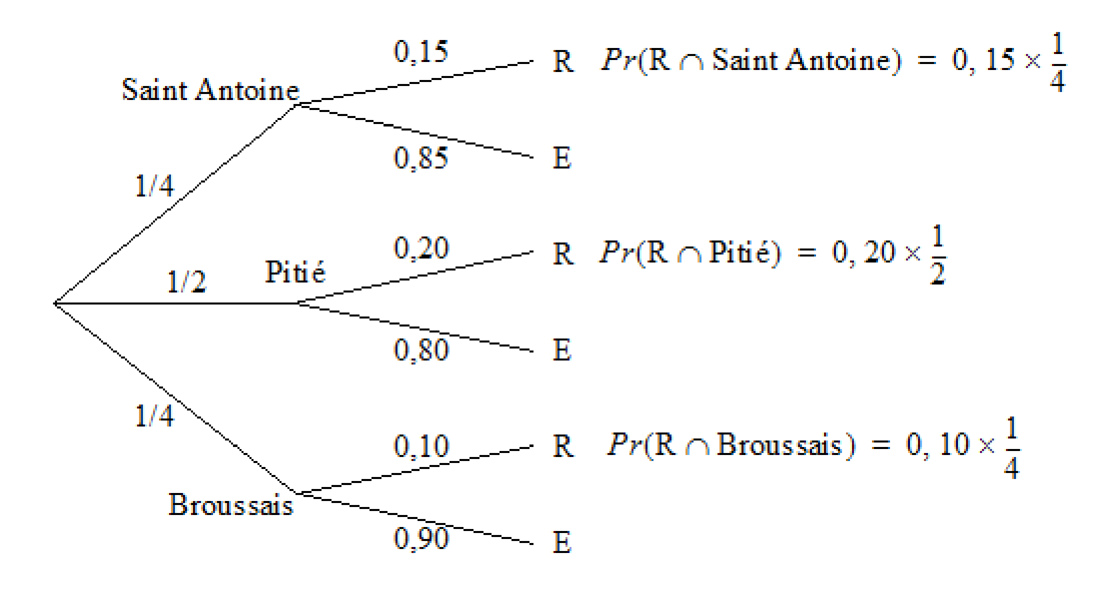
Exemple

Une boîte contient 10 articles dont 4 sont défectueux. On tire 3 objets de cette boîte. Calculer la probabilité pour que ces 3 objets soient défectueux.
Pr(1er défectueux) = 4/10
Pr(2ème défectueux / 1er défectueux) = 3/9
Pr(3ème défectueux / 1er et 2ème défectueux) = 2/8
Pr(1er et 2ème et 3ème défectueux) = 4/10×3/9×2/8 = 1/30.

4.3 Diagramme en arbre

On considère une séquence finie d'expériences dont chacune d'entre elles a un nombre fini de résultats possibles. Les probabilités associées aux résultats possibles d'une expérience dépendent du résultat de l'expérience précédente ; il s'agit de probabilités conditionnelles. Pour représenter cette séquence, on utilise une représentation « en arbre », le théorème précédent permettant de calculer la probabilité de chaque feuille de l'arbre.

Exemple : On sait que les taux de réussite au concours dans les trois CHU Pitié, Saint Antoine et Broussais (l'université Pierre et Marie Curie a longtemps comporté ces 3 CHU) étaient respectivement (données arbitraires) de 0,20 ; 0,15 ; et 0,10
(0,20 = *Pr*(Réussite/Pitié)) ; on sait que 1/4 des étudiants de Paris VI étaient à Saint Antoine, 1/4 à Broussais et 1/2 à la Pitié. Quelle était la probabilité qu'un étudiant de Paris VI soit reçu au concours ?



R signifie réussite et E échec.
$Pr(\text{R}) = Pr(\text{R} \cap \text{Saint Antoine}) + Pr(\text{R} \cap \text{Pitié}) + Pr(\text{R} \cap \text{Broussais})$
Pr(R) = 0,15×1/4 + 0,20×1/2 + 0,10×1/4 = 0,1625
La probabilité qu'un chemin particulier de l'arbre se réalise est, d'après le théorème de la multiplication, le produit des probabilités de chaque branche du chemin.
Les chemins s'excluant mutuellement, la probabilité d'être reçu est égale à la somme des probabilités d'être reçu pour tout chemin aboutissant à un état R (reçu).

4.4 - Théorème de BayesSections

En reprenant l'équation 2 (section 4.2), on obtient la formule de Bayes :

$$Pr(B/A) = \frac{Pr(A/B)Pr(B)}{Pr(A)}$$

Equation 3 : formule de Bayes

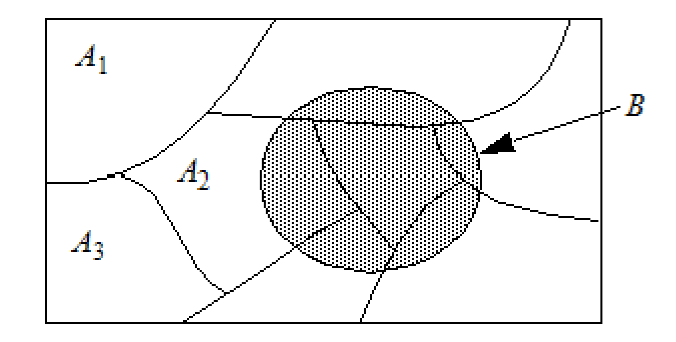
Le théorème est une forme développée de cette formule que nous introduisons maintenant.

Considérons des événements A_1, ..., A_n tels qu'ils forment une **partition** de l'ensemble fondamental *E*.

Par définition, les A_i s'excluent mutuellement et leur union est *E* :

$$\forall(i \neq j),(A_i \cap A_j = \varnothing) \; ; \; \bigcup_{i=1}^{n} A_i = E$$

Soit B un événement quelconque



De $E = A_1 \cup A_2 \cup \ldots \cup A_n$ et de $B \cap E = B$, on tire $B = B \cap (A_1 \cup A_2 \cup \ldots \cup A_n)$.

Soit, par distributivité, $B = (B \cap A_1) \cup (B \cap A_2) \cup \ldots \cup (B \cap A_n)$.

En remarquant que les $B \cap A_i$ sont exclusifs, puisque les A_i le sont, et en appliquant la 3ème règle du calcul des probabilités on obtient la formule dite des « probabilités totales » :

$$Pr(B) = Pr(B \cap A_1) + Pr(B \cap A_2) + \ldots + Pr(B \cap A_n)$$

Equation 4 : probabilités totales

En appliquant le théorème de la multiplication :

$$Pr(B) = Pr(B/A_1)Pr(A_1) + Pr(B/A_2)Pr(A_2) + \ldots + Pr(B/A_n)Pr(A_n)$$

Or, par la forme simple du théorème de Bayes, on a $Pr(A_i/B) = \dfrac{Pr(B/A_i)Pr(A_i)}{Pr(B)}$

D'où le théorème de Bayes :

$$Pr(A_i/B) = \frac{Pr(B/A_i)Pr(A_i)}{Pr(B/A_1)Pr(A_1) + Pr(B/A_2)Pr(A_2) + \ldots + Pr(B/A_n)Pr(A_n)}$$

Equation 5 : théorème de Bayes

Exemple 1

Reprenons l'exemple des résultats au concours des étudiants de Paris VI.
Comme précédemment, soit R l'événement « un étudiant de Paris VI est reçu ». On a, en notant C_1, C_2, C_3 les 3 anciens CHU Saint Antoine, Pitié et Broussais respectivement :
$Pr(\mathrm{R}) = Pr(\mathrm{R}/C_1)Pr(C_1) + Pr(\mathrm{R}/C_2)Pr(C_2) + Pr(\mathrm{R}/C_3)Pr(C_3)$
[noter que c'est la même chose que la somme des probabilités des chemins de l'arbre, qui conduisent à un succès]
Le théorème de Bayes permet de répondre à la question duale. Au lieu de chercher la probabilité d'obtenir un étudiant reçu sachant qu'il venait d'un CHU donné, on cherche la probabilité qu'un étudiant ait été inscrit à un CHU donné sachant qu'il a été reçu (probabilité des causes).

Calculons la probabilité qu'un étudiant reçu soit issu du CHU Pitié-Salpêtrière.

$$Pr(C_2/\mathrm{R}) = \frac{Pr(\mathrm{R}/C_2)Pr(C_2)}{Pr(\mathrm{R}/C_1)Pr(C_1) + Pr(\mathrm{R}/C_2)Pr(C_2) + Pr(\mathrm{R}/C_3)Pr(C_3)}$$

Avec $Pr(C_1) = 0{,}25$; $Pr(C_2) = 0{,}50$; $Pr(C_3) = 0{,}25$;
et $Pr(\mathrm{R}/C_1) = 0{,}15$; $Pr(\mathrm{R}/C_2) = 0{,}20$; $Pr(\mathrm{R}/C_3) = 0{,}10$.

D'où $$Pr(C_2/\mathrm{R}) = \frac{0,20 \times 0,50}{0,15 \times 0,25 + 0,20 \times 0,50 + 0,10 \times 0,25} = 0,61$$

Ce qui signifie que, dans ce cas, la probabilité qu'un étudiant appartienne à C_2, s'il est reçu, est plus grande que si l'on ne sait rien (probabilité a priori $Pr(C_2) = 0{,}50$). Cette façon de calculer les probabilités des causes connaissant les effets est essentielle en médecine. En effet, le problème du diagnostic peut être posé en ces termes.

Exemple 2

Considérons, pour illustrer notre propos, le problème du diagnostic d'une douleur aiguë de l'abdomen. Il s'agit d'un patient arrivant aux urgences pour un « mal au ventre ».
Si l'on ne sait rien d'autre sur le patient (on n'a pas fait d'examen clinique ou complémentaire), on ne connaît que les probabilités d'avoir tel ou tel diagnostic si on observe une douleur.
Soient D_1, D_2 et D_3 les 3 diagnostics principaux (il y en a en fait au moins une douzaine) et exclusifs ; par exemple D_1 = appendicite, D_2 = perforation d'ulcère, D_3 = autres diagnostics.
Soit un signe s_1 pour lequel on connaît $Pr(s_1/D_1)$, $Pr(s_1/D_2)$, et $Pr(s_1/D_3)$.
Par exemple, s_1 serait « présence d'une fièvre ≥ 38,5?C » ; $Pr(s_1/D_1) = 0{,}90$; $Pr(s_1/D_2) = 0{,}30$; et$Pr(s_1/D_3) = 0{,}10$.
Ces probabilités peuvent être estimées sur une population de patients en dénombrant le nombre de sujets ayant le diagnostic D_1 et présentant le signe s_1. De même, on peut connaître $Pr(D_1)$, $Pr(D_2)$ et $Pr(D_3)$.
Le problème diagnostique se pose comme celui de choisir par exemple le diagnostic le plus probable connaissant le signe s_1 ; pour ce faire, on calcule $Pr(D_1/s_1)$, $Pr(D_2/s_1)$, $Pr(D_3/s_1)$ et on retient le diagnostic qui a la plus grande probabilité : c'est l'application de l'approche bayesienne au problème de l'aide au diagnostic.

4.5 Indépendance entre événements

On dit que deux événements A et B sont indépendants si la probabilité pour que A soit réalisé n'est pas modifiée par le fait que B se soit produit. On traduit cela par $Pr(A / B) = Pr(A)$.

D'après la définition d'une probabilité conditionnelle, $Pr(A/B) = \frac{Pr(A \cap B)}{Pr(B)}$, on tire la définition :

A et B sont indépendants si et seulement si $Pr(A \cap B) = Pr(A)Pr(B)$.

La symétrie de cette définition implique qu'on a aussi bien $Pr(A / B) = Pr(A)$ (A est indépendant de B) que $Pr(B /A) = Pr(B)$ (B est indépendant de A) : l'apparition d'un des deux événements n'influe pas sur l'apparition de l'autre.

Note

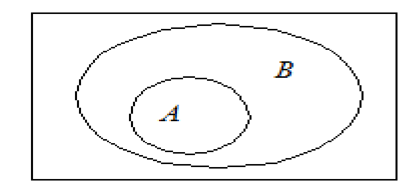
Ce qui est défini précédemment est l'indépendance de deux événements. Si on considère maintenant 3 événements *A*, *B*, *C*, on dira que ces 3 événements sont indépendants :

1. s'ils sont indépendants 2 à 2 : *A* indépendant de *B* ; *A* indépendant de *C* ; et *B* indépendant de *C*
2. et si $Pr(A \cap B \cap C) = Pr(A)Pr(B)Pr(C)$. Cette condition n'est pas une conséquence des précédentes.

4.6 Indépendance, inclusion et exclusion de deux événements

Considérons deux événements *A* et *B*.

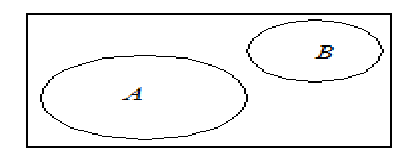
1. Si $A \subset B$ (*A* est inclus dans *B*) : si *A* est réalisé, alors *B* aussi.



2. Alors $Pr(A \cap B) = Pr(A)$.

   D'où $Pr(B / A) = \dfrac{Pr(A \cap B)}{Pr(A)} = 1$ et $Pr(A / B) = \dfrac{Pr(A \cap B)}{Pr(B)} = \dfrac{Pr(A)}{Pr(B)}$.

   A et *B* ne sont **pas indépendants**.
3. Si $A \cap B = \varnothing$ (*A* et *B* sont exclusifs) : si *A* est réalisé, *B* ne peut pas l'être.



4. Alors $Pr(A \cap B) = Pr(\varnothing) = 0$.

   D'où $Pr(A / B) = \dfrac{Pr(A \cap B)}{Pr(B)} = \dfrac{0}{Pr(B)} = 0$.

   De même *A* et *B* ne sont **pas indépendants**.

Chapitre 5 : Evaluation de l'intérêt diagnostique des informations médicales

5.1 - Introduction

La tâche essentielle des médecins est de traiter les patients. Pour prescrire un traitement, il faut savoir, plus ou moins précisément selon les cas, ce dont souffre le malade. Pour résumer en un seul terme un processus physiopathologique complexe, les médecins ont créé des concepts : les diagnostics.

La recherche « du » diagnostic est donc la première étape de la consultation clinique. Pour parvenir au diagnostic, le médecin accumule des informations, dont certaines lui sont spontanément livrées par le patient (le motif de la consultation, les symptômes), d'autres doivent être recherchées mais sont disponibles immédiatement (les signes physiques), d'autres enfin sont d'obtention plus ou moins difficile et coûteuse (les résultats d'examens complémentaires). De nouvelles procédures diagnostiques apparaissent fréquemment : on a vu, par exemple, l'apparition des échographies, de la tomodensitométrie (scanner), de l'IRM, pour ne citer que le domaine de l'imagerie. Il n'est bien sûr pas question d'effectuer tous les examens complémentaires sur tous les malades : il faut donc préciser les indications de ces examens, ce qui repose sur l'évaluation de leur intérêt diagnostique. Avant d'aborder la méthodologie de l'évaluation, nous reviendrons sur certains concepts utilisés dans ce paragraphe.

5.1.1 Le diagnostic

On peut définir un diagnostic comme un concept résumant l'état d'un individu. Le terme de « diagnostic » est donc beaucoup moins précis qu'on pourrait le penser à première vue : on peut en général fournir plusieurs diagnostics pour un même état physiopathologique, les termes diagnostiques utilisés dépendant de l'aspect privilégié. Parmi ces aspects, on peut citer :

- la symptomatologie
- la physiopathologie et l'étiologie
- la conduite thérapeutique

En pratique, la précision du diagnostic dépendra souvent des possibilités thérapeutiques : par exemple, on ne recherchera pas, en général, le virus responsable d'un syndrome grippal, surtout si on s'attend à ce que la maladie guérisse spontanément.

D'un point de vue statistique, le diagnostic sera souvent considéré comme une variable aléatoire binaire : le patient souffre ou ne souffre pas de l'affection considérée, ou, exprimé autrement, le diagnostic est vrai ou faux chez ce patient. Les valeurs possibles de la variable seront notées M et $\overline{M}$ (maladie présente ou absente), ou D et $\overline{D}$ (diagnostic vrai ou faux).

5.1.2 Les informations médicales

On divise l'ensemble des informations médicales en signes cliniques et signes complémentaires. Les signes cliniques sont divisés en signes fonctionnels ou symptômes, décrits par le malade (spontanément ou par l'interrogatoire) et signes physiques, recherchés par le médecin. Les signes complémentaires peuvent être biologiques ou radiologiques. Leur intérêt peut être :

- diagnostique (caractère malin ou bénin d'une tumeur)
- thérapeutique (localisation précise d'une tumeur)
- pronostique (extension ganglionnaire)

D'un point de vue statistique, ces signes peuvent être représentés par des variables binaires (présence ou absence d'un nodule sur une image) ou continues (cholestérolémie).

Nous considérons ici le seul cas d'un signe binaire, présent (noté S) ou absent (noté $\bar{S}$). Dans la suite, on considère que la présence du signe est évocateur de la maladie M.

Si l'information est de type continu, on se ramène au cas binaire par l'introduction d'un seuil : d'un côté du seuil, les valeurs sont dites normales, et le signe binaire est absent ; de l'autre côté du seuil, les valeurs sont dites pathologiques, et le signe binaire est présent.

5.1.3 Situation expérimentale et estimation

Quand on cherche à évaluer l'intérêt diagnostique d'un signe pour une affection, on recherche le signe chez des individus présentant ou non l'affection considérée. Deux situations expérimentales sont à envisager :

- **un échantillon** représentatif d'une population est constitué. On pourra estimer, à partir de cet échantillon, toutes les probabilités d'événements par les fréquences observées correspondantes (cette manière de faire sera revue plus tard, référence croisée non définie) ;
- **deux échantillons** sont constitués, l'un représentatif des individus pour lesquels le diagnostic est vrai, l'autre représentatif des individus pour lesquels il est faux. Cette manière de procéder est souvent la seule possible en pratique, surtout quand la maladie considérée est rare. Il faut remarquer, cependant, qu'on ne peut plus estimer n'importe quelle probabilité par la fréquence observée correspondante ; ce point sera développé plus loin dans ce chapitre.

Remarque : nous utilisons actuellement le mot *estimation* dans le sens d'*approximation* de la vraie valeur. Nous donnerons des définitions plus rigoureuses dans le chapitre 10.

5.2 - Les paramètres de l'évaluation

5.2.1 Sensibilité et spécificité

La **sensibilité** d'un signe pour une maladie est la probabilité que le signe soit présent si le sujet est atteint de la maladie considérée.

Il s'agit donc de la probabilité conditionnelle qu'on peut noter :

Sensibilité = Se =$Pr(S / M)$

Un test diagnostic est donc d'autant plus sensible que les sujets atteints de la maladie présentent plus souvent le signe S.

La **spécificité** d'un signe pour une maladie est la probabilité que le signe soit absent si le sujet n'est pas atteint de la maladie.

De manière similaire, on a :

Spécificité = Sp = $Pr(\bar{S}/\overline{M})$

Un test diagnostic est donc d'autant plus spécifique que les sujets indemnes de la maladie présentent moins souvent le signe S.

Pour un examen « parfait », c'est-à-dire n'effectuant aucune erreur, les valeurs de la sensibilité et de la spécificité sont égales à 1.

Si la présence du signe est définie par un « seuil de positivité », on observe que ces deux paramètres varient en sens inverse lorsqu'on fait varier ce seuil. Ceci explique qu'un seul de ces deux paramètres ne suffise pas à évaluer un examen. Supposons par exemple qu'on s'intéresse au signe température vis à vis de la grippe. On considère que le signe est présent si la température dépasse un certain seuil, par exemple 39?C. Si on augmente le seuil pour le porter à 40?C, la probabilité de dépasser le seuil (chez les sujets grippés) va diminuer, donc la sensibilité diminue. En revanche, la probabilité d'être en dessous du seuil (chez les sujets non grippés) va augmenter, donc la spécificité augmente.

Un test diagnostique de bonne sensibilité conduit à un résultat positif chez presque tous les malades. Il est donc utilisable pour un dépistage. Si le test possède une bonne spécificité, il conduit à un résultat négatif chez presque tous les non-malades. Il pourrait donc être utilisé en tant qu'examen de confirmation du diagnostic.

Ces considérations sont bien sûr schématiques, d'autres éléments intervenant dans l'évaluation, comme la fréquence de la maladie (prévalence), les risques liés à la maladie, à l'examen, l'existence et les performances d'autres examens concurrents...

5.2.2 Valeurs prédictives

En pratique, quand un médecin reçoit le résultat d'un examen complémentaire, positif ou négatif, il ne sait pas si le patient souffre de l'affection qu'il cherche à diagnostiquer ou non, et les probabilités qui l'intéressent s'expriment de la manière suivante : quelle est la probabilité de présence de la maladie M chez ce patient, sachant que l'examen a donné un résultat positif (ou négatif) ? Ces probabilités sont appelées valeurs prédictives. Plus précisément, on a :

- la **valeur prédictive positive** d'un signe pour une maladie est la probabilité que le sujet soit atteint de la maladie si le signe est présent ;
- la **valeur prédictive négative** d'un signe pour une maladie est la probabilité que le sujet soit indemne de la maladie si le signe est absent.

On peut noter ces paramètres :

$$\text{VPP} = Pr(M/S)$$

$$\text{VPN} = Pr(\bar{M}/\bar{S})$$

Comme les sensibilités et spécificité, les valeurs prédictives positive et négative varient en sens inverse, et doivent donc être considérées simultanément.

Les valeurs prédictives peuvent s'exprimer en fonction du couple sensibilité - spécificité, et de la fréquence de la maladie dans la population (cette probabilité $Pr(M)$ s'appelle la **prévalence** de la maladie). Il suffit d'utiliser le théorème de Bayes :

$$\begin{aligned}\text{VPP} = Pr(M/S) &= \frac{Pr(S/M)Pr(M)}{Pr(S/M)P(M) + Pr(S/\bar{M})Pr(\bar{M})} \\ &= \frac{\text{Se} \times Pr(M)}{\text{Se} \times Pr(M) + (1 - \text{Sp}) \times (1 - Pr(M))}\end{aligned}$$

$$\begin{aligned}\text{VPN} = Pr(\bar{M}/\bar{S}) &= \frac{Pr(\bar{S}/\bar{M})Pr(\bar{M})}{Pr(\bar{S}/M)P(M) + Pr(\bar{S}/\bar{M})Pr(\bar{M})} \\ &= \frac{\text{Sp} \times (1 - Pr(M))}{(1 - \text{Se}) \times Pr(M) + \text{Sp} \times (1 - Pr(M))}\end{aligned}$$

5.2.3 Comparaison des deux couples de paramètres

En situation clinique, on a vu que les valeurs prédictives correspondent aux préoccupations des médecins, et elles pourraient sembler les « meilleurs » paramètres d'évaluation. Pourtant, en réalité, c'est la sensibilité et la spécificité qui sont le plus souvent utilisées pour évaluer les examens complémentaires. La raison en est la suivante :

la sensibilité d'un examen pour une affection repose sur la définition de la population des « malades », et est donc caractéristique de la maladie et du signe. En particulier, elle n'est pas susceptible de varier d'un centre à l'autre (d'un service hospitalier spécialisé à une

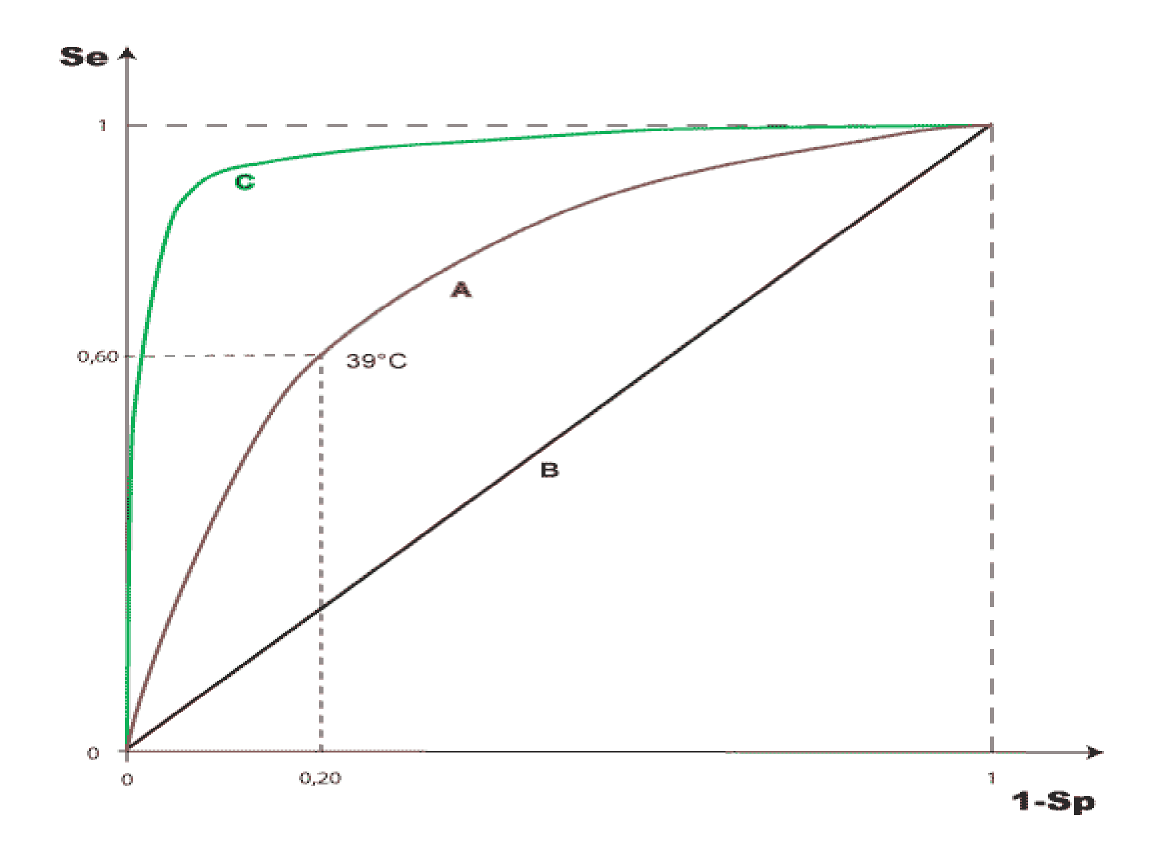
consultation de médecin généraliste, par exemple). Le même raisonnement peut s'appliquer à la spécificité, si on considère qu'elle repose aussi sur la définition de la maladie.

Les valeurs prédictives, au contraire, sont fonctions des proportions respectives de malades et de non-malades dans la population (de la prévalence de la maladie). Or ces proportions sont dépendantes des centres considérés ; les valeurs prédictives des examens varient donc d'un centre à l'autre pour une même maladie, ce qui explique qu'elles sont moins utilisées comme paramètre d'évaluation, même si elles sont intéressantes à connaître pour un centre donné.

5.2.4 Choix d'un seuil : courbes ROC

Lorsqu'un examen fournit des résultats de type continu, il faut déterminer le meilleur seuil entre les valeurs pathologiques et les valeurs normales. L'idéal serait d'obtenir une sensibilité et une spécificité égales à 1. Ce n'est généralement pas possible, et il faut tenter d'obtenir les plus fortes valeurs pour ces deux paramètres, sachant qu'ils varient en sens inverse.

On s'aide pour ce choix d'un outil graphique, la courbe ROC (*Receiver Operating Characteristics*). Une courbe ROC est le tracé des valeurs de la sensibilité Se en fonction de 1-Sp.



Cet exemple (tiré du livre de A.J. Valleron) montre 3 courbes ROC correspondant à 3 examens différents.

La courbe A est celle obtenue pour l'exemple précédent de la température et de la grippe. Le point de la courbe le plus proche du coin supérieur gauche du carré contenant la courbe (ici Se = 0,65, Sp = 0,75, et température = 39?C)) est celui qui permet d'obtenir un bon compromis entre sensibilité et spécificité (le coin supérieur gauche correspond à Se = Sp = 1). En réalité, on ne choisira pas toujours ce point, car il faut aussi tenir compte des coûts des erreurs diagnostiques : il peut par exemple être beaucoup plus grave de ne pas détecter une maladie, que de traiter à tort.

La courbe B correspond à un examen qui n'apporte rien au diagnostic, puisque les variables signe et maladie sont ici indépendantes : $\text{Se} = Pr(S/M) = 1 - \text{Sp} = Pr(S/\overline{M})$

La courbe C correspond à un bon critère diagnostic pour lequel on peut obtenir simultanément des valeurs élevées de sensibilité et de spécificité.

5.3 - Estimation des paramètres de l'évaluation

5.3.1 Un échantillon représentatif

5.3.1.1 Les données

Quand on a un échantillon représentatif d'une population, on peut résumer les données de l'expérience par un tableau de contingence 2×2, sur lequel sont indiqués les effectifs suivants :

- VP (Vrais Positifs) : ce sont les individus malades (*M*) et chez lesquels le signe est présent {*S*} ;
- FP (Faux Positifs) : la maladie est absente {$\overline{M}$} et le signe est présent {*S*} ;
- FN (Faux Négatifs) : la maladie est présente {*M*) et le signe est absent {$\bar{S}$} ;
- VN (Vrais Négatifs) : la maladie est absente {$\overline{M}$} et le signe est absent {$\bar{S}$}.

Tableau 1

	M	$\overline{M}$
S	VP	FP
$\bar{S}$	FN	VN

5.3.1.2 Estimation de la sensibilité et de la spécificité

Par définition, sensibilité = Se = $Pr(S / M)$

On estime cette probabilité conditionnelle par le rapport des effectifs correspondants sur le tableau de contingence observé :

$$Se \approx \frac{VP}{VP + FN}$$

Note : On notera de manière identique, suivant un usage établi, les paramètres vrais, qui sont des probabilités conditionnelles, et leurs estimations, qui sont des rapports d'effectifs observés.

$$\text{Spécificité} = Sp = Pr(\bar{S}/\bar{M}) \approx \frac{VN}{VN + FP}$$

Par exemple, calculons les estimateurs de ces paramètres dans le cas où on cherche à diagnostiquer un diabète à partir d'un signe de la forme « la glycémie mesurée à jeun est supérieure à ... »). Pour deux seuils donnés S_1 et S_2, on obtient les tableaux de contingence ci-dessous :

a. Seuil S_1

Tableau 2

	M	$\bar{M}$
S	90	200
$\bar{S}$	10	300

b. Seuil S_2

Tableau 3

	M	$\bar{M}$
S	50	25
$\bar{S}$	50	475

On peut estimer les sensibilités et spécificités correspondant aux deux seuils par :

$Se_1 \approx 90 / 100 = 0,90$; $Sp_1 \approx 300 / 500 = 0,60$

$Se_2 \approx 50 / 100 = 0,50$; $Sp_2 \approx 475 / 500 = 0,95$.

On retrouve ici le fait que sensibilité et spécificité varient en sens inverse.

On constate d'autre part que le seuil S_1 correspond à une bonne sensibilité (l'examen est positif chez 90 % des malades), mais à une spécificité médiocre (l'examen est positif chez 40 % des « non-malades ») ; il peut donc être utilisé pour un examen de dépistage, le diagnostic devant être confirmé ultérieurement par un examen plus spécifique.

Le seuil S_2, en revanche, induit un test d'une sensibilité qui pourrait être jugée trop faible pour un examen de dépistage. En revanche, sa spécificité peut être acceptable pour un examen de confirmation.

5.3.1.3 Estimation des valeurs prédictives

Les estimations s'obtiennent à partir du même tableau des données :

$$\text{VPP} = Pr(M/S) \approx \frac{\text{VP}}{\text{VP}+\text{FP}}$$

$$\text{VPN} = Pr(\bar{M}/\bar{S}) \approx \frac{\text{VN}}{\text{VN}+\text{FN}}$$

Par exemple, pour les tableaux de contingence vus ci-dessus, on a :

$VPP_1 \approx 90 / 290 = 0,31$; $VPN_1 \approx 300 / 310 = 0,97$

$VPP_2 \approx 50 / 75 = 0, 67$; $VPN_2 \approx 475 / 525 = 0,90$

Ces résultats peuvent s'interpréter ainsi : en affirmant le diagnostic sur la base de la positivité de l'examen, on se trompe dans 69 % des cas avec le seuil S_1 et 33 % des cas avec le seuil S_2 ; et en éliminant le diagnostic en constatant la négativité de l'examen, on se trompe dans 3 % des cas avec le seuil S_1 et 10 % des cas avec le seuil S_2.

5.3.2 Deux échantillons représentatifs

L'inconvénient du schéma expérimental ci-dessus (un seul échantillon) est que, si la maladie est peu fréquente ou rare, il faut constituer un échantillon de très grande taille pour obtenir un nombre suffisant de malades. Les non-malades, au contraire, seront « trop » nombreux. C'est pourquoi on constituera souvent, en pratique, deux échantillons, un échantillon de malades et un échantillon de non-malades. On peut encore résumer les résultats par un tableau comme celui du tableau 1, mais ce tableau doit être interprété différemment, les proportions respectives des malades et non-malades ne correspondant plus à la réalité : le rapport entre le nombre de malades et le nombre de non-malades du tableau dépend des

tailles respectives choisies pour les deux échantillons, et n'a aucun lien avec la fréquence de la maladie dans la population (**la prévalence**).

On peut toujours estimer la sensibilité et la spécificité comme ci-dessus. En effet, la sensibilité par exemple est estimée uniquement à partir de VP et FN, donc de la répartition des malades entre ceux qui présentent le signe et les autres. Or l'échantillon des malades respecte cette répartition.

En revanche, l'estimation précédente des valeurs prédictives utilisait la répartition entre malades et non malades, que le tableau actuel ne représente pas correctement.

L'estimation des valeurs prédictives reste cependant possible à condition de connaître la prévalence de la maladie$Pr(M)$. On utilisera les formules introduites section 5.2.2 :

$$\text{VPP} = \frac{\text{Se} \times Pr(M)}{\text{Se} \times Pr(M) + (1 - \text{Sp}) \times (1 - Pr(M))}$$

$$\text{VPN} = \frac{\text{Sp} \times (1 - Pr(M))}{(1 - \text{Se}) \times Pr(M) + \text{Sp} \times (1 - Pr(M))}$$

On remplacera dans ces formules la sensibilité et la spécificité par leurs estimations.

5.3.1 Un échantillon représentatif

5.3.1.1 Les données

Quand on a un échantillon représentatif d'une population, on peut résumer les données de l'expérience par un tableau de contingence 2×2, sur lequel sont indiqués les effectifs suivants :

- VP (Vrais Positifs) : ce sont les individus malades (M) et chez lesquels le signe est présent {S} ;
- FP (Faux Positifs) : la maladie est absente {$\bar{M}$} et le signe est présent {S} ;
- FN (Faux Négatifs) : la maladie est présente {M) et le signe est absent {$\bar{S}$} ;
- VN (Vrais Négatifs) : la maladie est absente {$\bar{M}$} et le signe est absent {$\bar{S}$}.

Tableau 1

	M	$\bar{M}$
S	VP	FP
$\bar{S}$	FN	VN

5.3.1.2 Estimation de la sensibilité et de la spécificité

Par définition, sensibilité = Se = $Pr(S / M)$

On estime cette probabilité conditionnelle par le rapport des effectifs correspondants sur le tableau de contingence observé :

$$Se \approx \frac{VP}{VP + FN}$$

Note : On notera de manière identique, suivant un usage établi, les paramètres vrais, qui sont des probabilités conditionnelles, et leurs estimations, qui sont des rapports d'effectifs observés.

$$\text{Spécificité} = Sp = Pr(\bar{S}/\bar{M}) \approx \frac{VN}{VN + FP}$$

Par exemple, calculons les estimateurs de ces paramètres dans le cas où on cherche à diagnostiquer un diabète à partir d'un signe de la forme « la glycémie mesurée à jeun est supérieure à ... »). Pour deux seuils donnés S_1 et S_2, on obtient les tableaux de contingence ci-dessous :

a. Seuil S_1

Tableau 2

	M	$\bar{M}$
S	90	200
$\bar{S}$	10	300

b. Seuil S_2

Tableau 3

	M	$\bar{M}$
S	50	25
$\bar{S}$	50	475

5.3.1.3 Estimation des valeurs prédictives

Les estimations s'obtiennent à partir du même tableau des données :

$$\text{VPP} = Pr(M/S) \approx \frac{\text{VP}}{\text{VP} + \text{FP}}$$

$$\text{VPN} = Pr(\bar{M}/\bar{S}) \approx \frac{\text{VN}}{\text{VN} + \text{FN}}$$

Par exemple, pour les tableaux de contingence vus ci-dessus, on a :
$VPP_1 \approx 90 / 290 = 0{,}31$; $VPN_1 \approx 300 / 310 = 0{,}97$
$VPP_2 \approx 50 / 75 = 0{,} 67$; $VPN_2 \approx 475 / 525 = 0{,}90$
Ces résultats peuvent s'interpréter ainsi : en affirmant le diagnostic sur la base de la positivité de l'examen, on se trompe dans 69 % des cas avec le seuil S_1 et 33 % des cas avec le seuil S_2 ; et en éliminant le diagnostic en constatant la négativité de l'examen, on se trompe dans 3 % des cas avec le seuil S_1 et 10 % des cas avec le seuil S_2.

5.3.2 Deux échantillons représentatifs

L'inconvénient du schéma expérimental ci-dessus (un seul échantillon) est que, si la maladie est peu fréquente ou rare, il faut constituer un échantillon de très grande taille pour obtenir un nombre suffisant de malades. Les non-malades, au contraire, seront « trop » nombreux. C'est pourquoi on constituera souvent, en pratique, deux échantillons, un échantillon de malades et un échantillon de non-malades.

On peut encore résumer les résultats par un tableau comme celui du tableau 1, mais ce tableau doit être interprété différemment, les proportions respectives des malades et non-malades ne correspondant plus à la réalité : le rapport entre le nombre de malades et le nombre de non-malades du tableau dépend des tailles respectives choisies pour les deux échantillons, et n'a aucun lien avec la fréquence de la maladie dans la population (**la prévalence**).

On peut toujours estimer la sensibilité et la spécificité comme ci-dessus. En effet, la sensibilité par exemple est estimée uniquement à partir de VP et FN, donc de la répartition des malades entre ceux qui présentent le signe et les autres. Or l'échantillon des malades respecte cette répartition.

En revanche, l'estimation précédente des valeurs prédictives utilisait la répartition entre malades et non malades, que le tableau actuel ne représente pas correctement.

L'estimation des valeurs prédictives reste cependant possible à condition de connaître la prévalence de la maladie$Pr(M)$. On utilisera les formules introduites section 5.2.2 :

$$\text{VPP} = \frac{\text{Se} \times Pr(M)}{\text{Se} \times Pr(M) + (1 - \text{Sp}) \times (1 - Pr(M))}$$

$$\text{VPN} = \frac{\text{Sp} \times (1 - Pr(M))}{(1 - \text{Se}) \times Pr(M) + \text{Sp} \times (1 - Pr(M))}$$

On remplacera dans ces formules la sensibilité et la spécificité par leurs estimations.

Chapitre 6 : Variables aléatoires

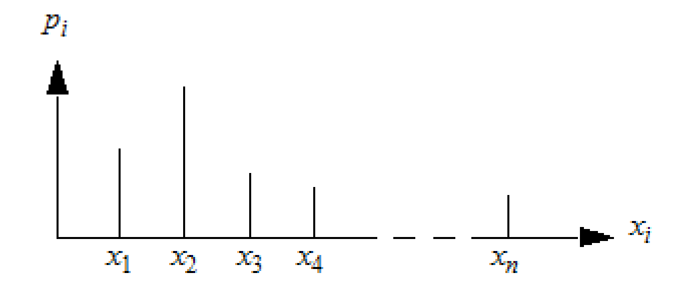
6.1 - Définition d'une variable aléatoire

Considérons un ensemble fondamental E correspondant à une certaine expérience. Les éléments de E, résultats possibles de l'expérience, ne sont généralement pas des nombres. Il est cependant utile de faire correspondre un nombre à chaque élément de E, en vue de faire ensuite des calculs. Pour un jet de dé, il semble naturel de faire correspondre à la face obtenue par le jet, le nombre de points qu'elle porte, mais ce n'est pas une obligation. Si on jette 2 dés, on s'intéressera par exemple à la somme des points obtenus. Pour une carte à jouer, il faut convenir d'une valeur pour chaque carte.
Une variable aléatoire X, sur un ensemble fondamental E, est une application de E dans $\Re$: à tout résultat possible de l'expérience (à tout élément de E), la variable aléatoire X fait correspondre un nombre.
Lorsque E est fini ou infini dénombrable, toute application de E dans $\Re$ est une variable aléatoire.
Lorsque E est non dénombrable, il existe certaines applications de E dans $\Re$ qui ne sont pas des variables aléatoires. En effet, la définition rigoureuse d'une variable aléatoire X impose que tout intervalle de $\Re$ soit l'image d'un événement de E par l'application X. Cette condition est vérifiée pour toute application X si E est fini ou dénombrable, puisque toute partie de E est un événement. Ce n'est plus vrai si E est non dénombrable. Heureusement, les applications choisies naturellement sont des variables aléatoires.
On parle de variable aléatoire **discrète** lorsque la variable est une application de E dans un sous-ensemble discret de$\Re$, le plus souvent **N** ou une partie de **N**. On parle sinon de variable aléatoire **continue**.
Pour un nombre réel a donné, l'événement constitué de tous les résultats ξ d'expérience tels que $X(\xi) = a$ est noté $[X(\xi) = a]$, ou, en abrégé, $X = a$.
Pour deux nombres réels a et b ($a \leq b$), l'événement constitué de tous les résultats ξ d'expérience tels que $a \leq X(\xi) \leq b$est noté $[a \leq X(\xi) \leq b]$ ou, en abrégé, $a \leq X \leq b$.
Si X et Y sont des variables aléatoires définies sur le même ensemble fondamental E, et si k est une constante, on peut montrer que les fonctions suivantes sont aussi des variables aléatoires :

$(X + Y)(\xi) = X(\xi) + Y(\xi)$ $\qquad$ $(X + k)(\xi) = X(\xi) + k$

$(kX)(\xi) = kX(\xi)$ $\qquad$ $(XY)(\xi) = X(\xi)\, Y(\xi)$

pour tout élément ξ de E.



6.3 Variables infinies dénombrables (hors programme)

Tout ce qui a été vu précédemment dans le cas où E est fini ($E = \{s_1, s_2, ..., s_n\}$) se généralise (nous ne verrons pas les démonstrations) au cas où E est infini dénombrable ; on aura par exemple

$$\mu_X = E(X) = \sum_{i=1}^{\infty} x_i p_i$$

La somme converge à l'infini vers $E(X)$, toutes les autres propriétés sont conservées, les sommes devenant des séries.

6.4 Variables aléatoires continues

La généralisation au continu est délicate et même difficile si on ne dispose pas d'outils mathématiques hors du champ de ce cours.

Nous nous contenterons de procéder par analogie avec le cas discret.

Une variable aléatoire X dont l'ensemble image $X(E)$ est un intervalle de $\mathfrak{R}$ est une variable aléatoire continue (continue par opposition à discrète, cf supra).

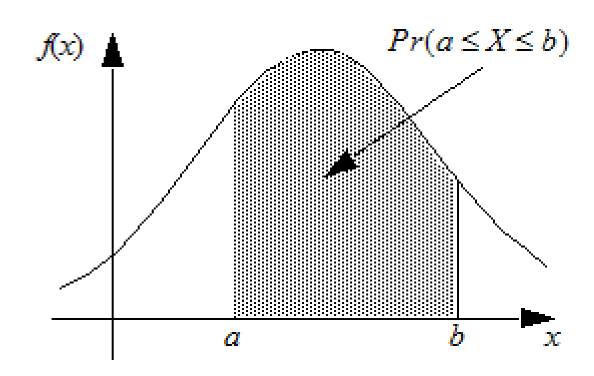
Rappelons que, par définition d'une variable aléatoire, $a \leq X \leq b$ est un événement de E dont la probabilité est bien définie.

On définit la loi de probabilité de X, ou distribution de X, à l'aide d'une fonction $f(x)$, appelée **densité de probabilité** de X, telle que

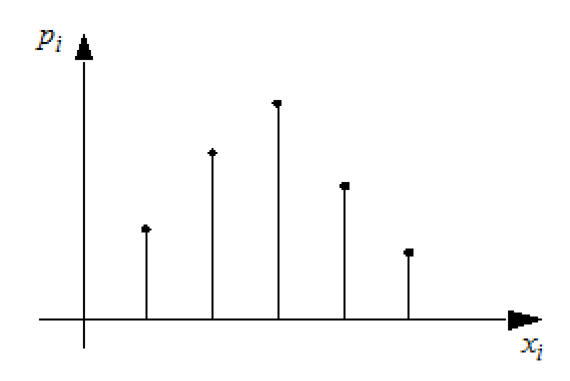
$$\int_a^b f(x)dx = Pr(a \leq X \leq b)$$

Remarques

1. Si f est donnée, la probabilité $Pr(a \leq X \leq b)$ est la surface sous la courbe entre a et b



2. Le passage du discret au continu transforme les sommes $\sum$ en intégrales $\int$ et p_i en $f(x)dx$.
   Ainsi, soit X une variable aléatoire discrète et p_i sa distribution

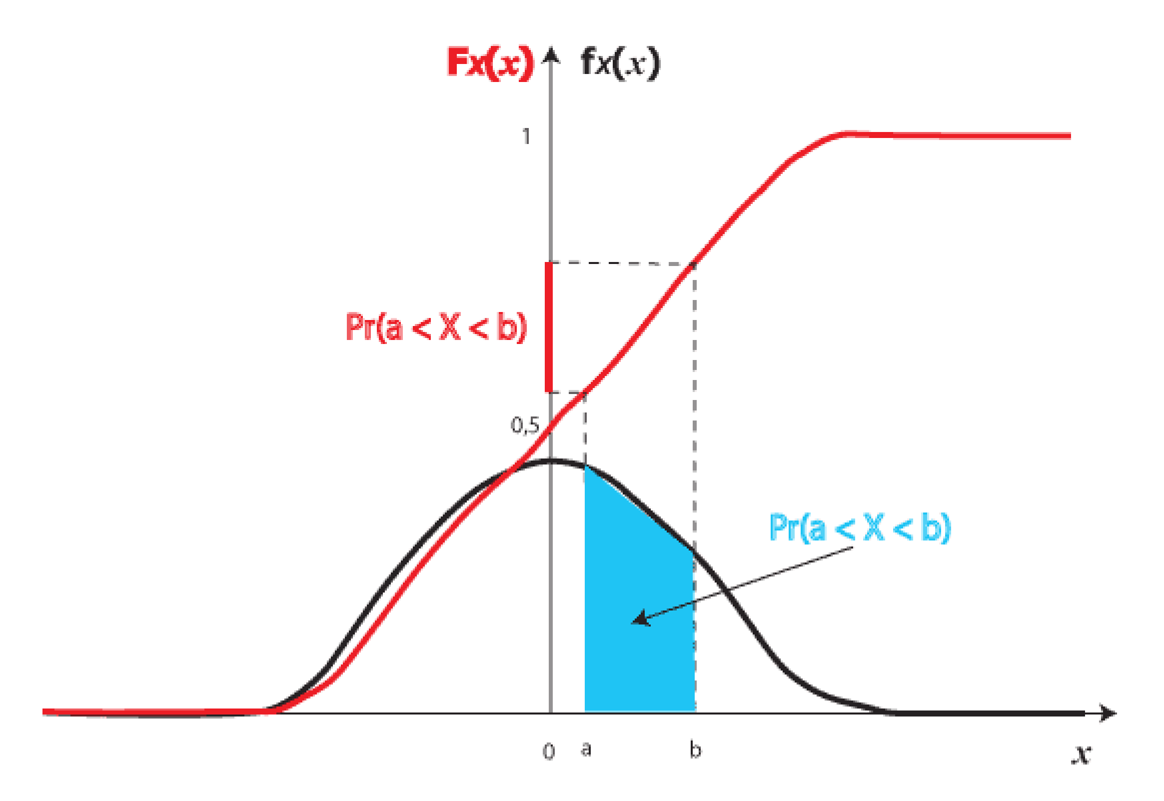


3. La formule $Pr(x_k \leq X \leq x_n) = \sum_{i=k}^{n} p_i$ est analogue à $Pr(a \leq X \leq b) = \int_a^b f(x)\,dx$

En utilisant cette analogie, on admettra les définitions suivantes pour une variable aléatoire *X*, continue, de distribution *f*(*x*) :

1. $f(x) \geq 0$ (analogue à $p_i \geq 0$)
2. $\int_{\Re} f(x)\,dx = 1$ (analogue à $\sum_i p_i = 1$)
3. $\mu_X = E(X) = \int_{\Re} x f(x)\,dx$ (analogue à $\sum_i x_i p_i$)
4. $\sigma_X^2 = var(X) = \int_{\Re} (x - \mu_X)^2 f(x)\,dx$ (analogue à $\sum_i (x_i - \mu_X)^2 p_i$)
5. $\sigma_X^2 = var(X) = \int_{\Re} x^2 f(x)\,dx - \mu_X^2$ (analogue à $\sum_i x_i^2 p_i - \mu_X^2$)
6. $\sigma(X) = \sigma_X = \sqrt{var(X)}$
7. $F(x) = Pr(X \leq x) = \int_{-\infty}^{x} f(\tau)\,d\tau$ (analogue à $\sum_{x_i \leq x} p_i$)

   Les propriétés de la fonction de répartition données section 6.2.6 sont conservées : fonction monotone croissante, partant de 0 pour $x \to -\infty$ et atteignant 1 pour $x \to +\infty$.
8. $Pr(a \leq X \leq b) = \int_a^b f(x)\,dx = F(b) - F(a)$



Cet exemple montre la densité de probabilité et la fonction de répartition d'une certaine variable aléatoire continue. La probabilité de l'intervalle [a b] est la surface sous la courbe de densité limitée par cet intervalle. C'est aussi la différence des hauteurs *F*(b)-*F*(a) si on utilise la fonction de répartition. Contrairement au cas des variables discrètes, la fonction de répartition est ici continue.Pour résumer l'analogie entre le cas discret et le cas continu, un point du domaine discret correspond à un intervalle dans le cas continu, la somme discrète correspond à l'intégrale.

6.5 Extension de la notion de variable aléatoire

Une variable aléatoire, telle qu'elle est définie dans ce chapitre, ne peut prendre que des valeurs numériques. Il est pourtant souvent pratique de s'intéresser directement aux résultats d'une expérience, qu'ils soient numériques ou non, c'est à dire d'éviter le codage numérique de ces résultats. Par abus de langage, dans la suite du cours, on pourra parler de variables aléatoires alors qu'il s'agit de résultats d'expérience.

Dans ce contexte, la classification antérieure des variables (discrètes ou continues) doit être étendue :

Variables quantitatives : Variables dont les valeurs sont numériques. C'est l'unique possibilité dans le cas de variables aléatoires au sens strict.

On distingue deux types de variables quantitatives :

- **variables discrètes**, dont les valeurs sont discrètes, en général des nombres entiers. Exemple : nombre d'étudiants dans un amphi.
- **variables continues**, pour lesquelles toutes les valeurs sont possibles, au moins sur un intervalle. Exemples : le poids ou la taille.

Variables qualitatives : Variables dont les valeurs ne sont pas numériques.

On en distingue deux types :

- **variables ordinales**, dont les valeurs peuvent être ordonnées. Exemple : intensité d'une douleur qui peut aller de *absente* à *très intense*.
- **variables catégorielles** ou **nominales**, dont les valeurs ne peuvent pas être ordonnées. Exemple : couleur des yeux.

Chapitre 7 : Exemples de distributions

7.1 - Lois discrètes

Les lois décrites ici ne concernent que des variables dont les valeurs sont des nombres entiers.

7.1.1 Loi de Bernoulli

On considère une expérience n'ayant que deux résultats possibles, par exemple succès et échec (ou présence et absence d'une certaine caractéristique). On introduit la variable aléatoire X qui associe la valeur 0 à l'échec (ou à l'absence de la caractéristique) et la valeur 1 au succès (ou à la présence de la caractéristique). Cette variable aléatoire est appelée variable de Bernoulli.

Distribution de *X*

Appelons Π la probabilité de l'événement succès :

$Pr(\{\text{succès}\}) = Pr(X = 1) = \Pi$

d'où

$Pr(\{\text{échec}\}) = Pr(X = 0) = 1 - \Pi$

Espérance de *X*

$$\mu_X = E(X) = \sum x_i Pr(X = x_i) = 1 \times Pr(X = 1) + 0 \times Pr(X = 0) = \Pi$$

Variance de *X*

$$\sigma_X^2 = var(X) = E[(X - \mu_X)^2] = E(X^2) - \mu_X^2$$

$$\sigma_X^2 = [1^2 \times Pr(X = 1) + 0^2 \times Pr(X = 0)] - \Pi^2$$

$$\sigma_X^2 = \Pi - \Pi^2 = \Pi(1 - \Pi)$$

7.1.2 Loi binomiale

Définition

Soient les épreuves répétées et indépendantes d'une même expérience de Bernoulli. Chaque expérience n'a que deux résultats possibles : succès ou échec. Comme précédemment, appelons Π la probabilité de l'événement élémentaire succès. A cette expérience multiple on associe une variable aléatoire X qui mesure le nombre de succès obtenus.

Distribution de *X*

On montre aisément que la probabilité d'avoir *k* succès lors de *n* épreuves répétées est

$$P(X = k \text{ pour } n \text{ essais}) = \frac{n!}{k!\,(n-k)!}\Pi^{k}(1-\Pi)^{n-k}$$

Rappel

$n! = 1 \times 2 \times \ldots \times n$ pour tout *n* entier positif

0! = 1 par définition

Remarques

a. La probabilité de n'avoir aucun succès au cours de *n* épreuves (*k* = 0) est $(1-\Pi)^n$; la probabilité d'avoir au moins un succès est donc $1 - (1-\Pi)^n$ (un succès ou plus)

b. $\frac{n!}{k!(n-k)!}$ est souvent noté $\binom{n}{k}$ ou C_n^k

Les $\binom{n}{k}$ s'appellent coefficients du binôme.

En effet ils interviennent dans le développement du binôme selon la formule

$$(a+b)^n = \sum_{r=0}^{n}\binom{n}{r}a^{n-r}b^{r}$$

Exercice :

utiliser cette formule pour vérifier que $(a+b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$

c. En appliquant la formule du binôme précédente on retrouve que la somme des probabilités pour toutes les valeurs de *X* est égale à 1 :

$$\sum_{k=0}^{n}\binom{n}{k}\Pi^{k}(1-\Pi)^{n-k} = [\Pi + (1-\Pi)]^{n} = 1^{n} = 1$$

Exemples

1. On jette 6 fois une pièce bien équilibrée ; on suppose que face est un succès. On a donc
   Π = 1/2 et *n* = 6

a. Probabilité que l'on ait exactement 2 faces

$$Pr(2 \text{ faces parmi } 6 \text{ jets}) = \frac{6!}{2!4!} \cdot \left(\frac{1}{2}\right)^2 \cdot \left(\frac{1}{2}\right)^4 = \frac{1 \times 2 \times 3 \times 4 \times 5 \times 6}{1 \times 2 \times 1 \times 2 \times 3 \times 4} \cdot \frac{1}{4} \cdot \frac{1}{16}$$

$$Pr(2 \text{ faces parmi } 6 \text{ jets}) = \frac{5 \times 6}{2 \times 4 \times 16} = \frac{15}{4 \times 16} = \frac{15}{64}$$

b. Probabilité d'avoir 4 faces ou plus (au moins 4 faces)
C'est aussi la probabilité d'avoir au plus 2 piles (0, 1 ou 2 piles)

$$p_4 = Pr(4 \text{ faces}) = \frac{6!}{2!4!} \cdot \left(\frac{1}{2}\right)^4 \cdot \left(\frac{1}{2}\right)^2 = \frac{1 \times 2 \times 3 \times 4 \times 5 \times 6}{1 \times 2 \times 1 \times 2 \times 3 \times 4} \cdot \frac{1}{16} \cdot \frac{1}{4} = \frac{15}{64}$$

$$p_5 = Pr(5 \text{ faces}) = \frac{6!}{1!5!} \cdot \left(\frac{1}{2}\right)^5 \cdot \frac{1}{2} = \frac{1 \times 2 \times 3 \times 4 \times 5 \times 6}{1 \times 2 \times 3 \times 4 \times 5} \cdot \frac{1}{32} \cdot \frac{1}{2} = \frac{6}{64}$$

$$p_6 = Pr(6 \text{ faces}) = \frac{6!}{6!} \cdot \left(\frac{1}{2}\right)^6 \cdot \left(\frac{1}{2}\right)^0 = \frac{1}{64}$$

$$Pr(\text{au moins } 4 \text{ faces}) = p_4 + p_5 + p_6 = \frac{15}{64} + \frac{6}{64} + \frac{1}{64} = \frac{11}{32}$$

2. On jette 7 fois un dé équilibré et on considère que tirer 5 ou 6 est un succès. Calculer

a. la probabilité pour qu'on ait 3 succès exactement

$$Pr(\text{succès}) = Pr(\{5, 6\}) = \frac{2}{6} = \frac{1}{3}$$

$$Pr(3 \text{ succès}) = \frac{7!}{3!4!}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^4 = \frac{560}{2187}$$

b. la probabilité de n'avoir aucun succès

$$Pr(\text{aucun succès}) = (1 - \Pi)^7 = \left(\frac{2}{3}\right)^7 = \frac{128}{2187}$$

Propriétés

La fonction de probabilité $Pr(X = k)$ dépend des 2 paramètres (ou constantes) n et Π.
C'est une distribution discrète qui prend les valeurs suivantes :

k	0	1	2	……..	n
$Pr(X = k)$	$(1-\Pi)^n$	$\binom{n}{1}(1-\Pi)^{n-1}\Pi$	$\binom{n}{2}(1-\Pi)^{n-2}\Pi^2$		Π^n

On dit que *X* est distribuée selon une loi binomiale B(*n*, Π). On peut montrer que

Distribution binomiale B(*n*, Π)	
Espérance	$\mu = n\Pi$
Variance	$\sigma^2 = n\Pi(1-\Pi)$
Ecart-type	$\sigma = \sqrt{n\Pi(1-\Pi)}$

7.1.3 Loi de Poisson

La loi de Poisson (due à Siméon Denis Poisson en 1837) est la loi du nombre d'événements observé pendant une période de temps donnée dans le cas où ces **événements** sont **indépendants et faiblement probables**. Elle peut s'appliquer au nombre d'accidents, à l'apparition d'anomalies diverses, à la gestion des files d'attentes, au nombre de colonies bactériennes dans une boîte de Pétri, etc.

Définition

Soit *X* la variable aléatoire représentant le nombre d'apparitions indépendantes d'un événement faiblement probable dans une population infinie. La probabilité d'avoir *k* apparitions de l'événement est

$$Pr(X=k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

Cette loi dépend d'un paramètre λ, nombre réel strictement positif.

Les nombres *k* possibles sont toutes les valeurs entières 0, 1, 2, etc. Cependant, lorsque k est suffisamment grand, la probabilité correspondante devient extrèmement faible.

Propriétés

- On peut montrer que

Loi de Poisson $\mathcal{P}(\lambda)$	
Espérance	$\mu = \lambda$
Variance	$\sigma^2 = \lambda$
Ecart-type	$\sigma = \sqrt{\lambda}$

- La démonstration utilise le fait que $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$
- Si deux variables aléatoires indépendantes X_1 et X_2 sont distribuées selon des lois de Poisson de paramètres λ_1 et λ_2, alors la variable X_1+X_2 est distribuée selon une loi de Poisson de paramètre $\lambda_1+\lambda_2$.

Remarques

Si on connaît la probabilité de n'observer aucun événement $Pr(X=0) = p$:

- D'après la formule, $p = e^{-\lambda}\frac{\lambda^0}{0!} = e^{-\lambda}$

  On en déduit :

  $\lambda = -\ln p$

- $Pr(X = 1) = e^{-\lambda}\frac{\lambda^1}{1!} = p\lambda$,

  $Pr(X = 2) = e^{-\lambda}\frac{\lambda^2}{2!} = Pr(X = 1)\frac{\lambda}{2}$,

  $Pr(X = 3) = e^{-\lambda}\frac{\lambda^3}{3!} = Pr(X = 2)\frac{\lambda}{3}$,

  ……

  $Pr(X = k) = Pr(X = k-1)\frac{\lambda}{k}$

  On peut ainsi calculer facilement de proche en proche les probabilités des diverses valeurs de *k*.

Lien avec la loi binomiale

Si une variable aléatoire *X* est distribuée selon une loi binomiale B(*n*, Π), on montre que si Π est petit (en pratique inférieur à 0,1) et *n* assez grand (supérieur à 50), la loi binomiale peut être approximée par une loi de Poisson de paramètre $\lambda = n\Pi$.
Les calculs sont plus simples avec la loi de Poisson qu'avec la binomiale.
Notons que puisque *X* est distribuée selon une loi binomiale, ses valeurs possibles ne peuvent dépasser *n*, alors que l'approximation par la loi de Poisson autorise des valeurs supérieures. Cependant le calcul fournit des probabilités très faibles pour ces valeurs aberrantes.

7.2 - Lois continues

7.2.1 - Loi normale

7.2.1.1 Définition

La distribution normale, ou de Laplace-Gauss, appelée aussi gaussienne, est une distribution continue qui dépend de deux paramètres μ et σ. On la note N(μ, σ^2). Le paramètre μ peut être quelconque mais σ est positif. Cette distribution est définie par :

$$f(x;\ \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

C'est une des lois les plus importantes, sinon la plus importante comme vous le verrez à l'occasion du théorème central limite.

7.2.1.2 Propriétés

Allure de la courbe

La loi normale, notée N(μ, σ^2), est symétrique par rapport à la droite d'abscisse μ.

Exemples :

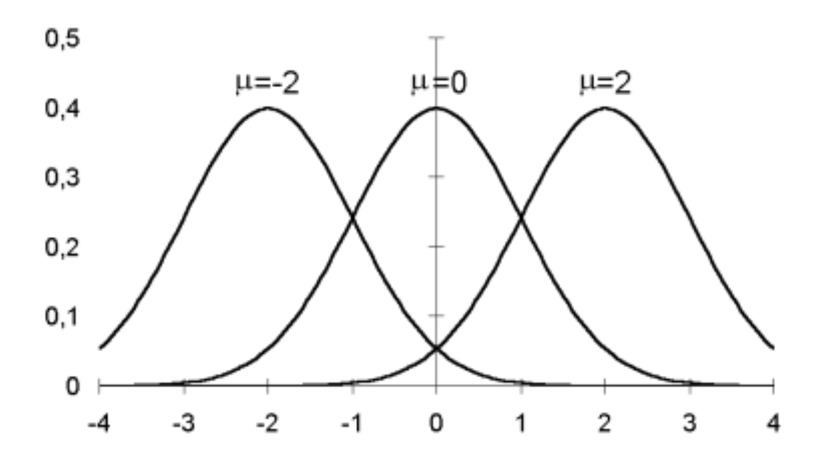


Figure 2 : N(μ, 1) pour les valeurs de μ -2 ; 0 et 2

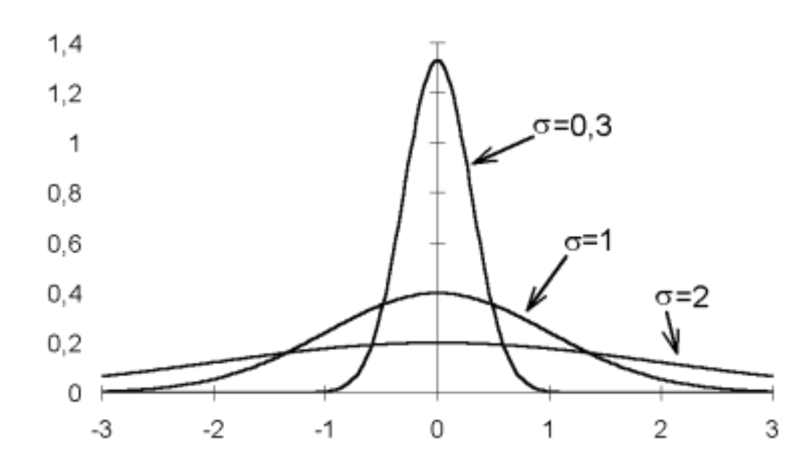


Figure 3 : N(0, σ^2) pour les valeurs de σ 0,3 ; 1 et 2

Caractéristiques

Loi normale N(μ, σ^2)	
Espérance	μ
Variance	σ^2
Ecart-type	σ

La distribution normale centrée réduite

On dit que la distribution est centrée si son espérance μ est nulle ; elle est dite réduite si sa variance σ^2(et son écart-type σ) est égale à 1. La distribution normale centrée réduite N(0, 1) est donc définie par la formule

$$f(t; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$$

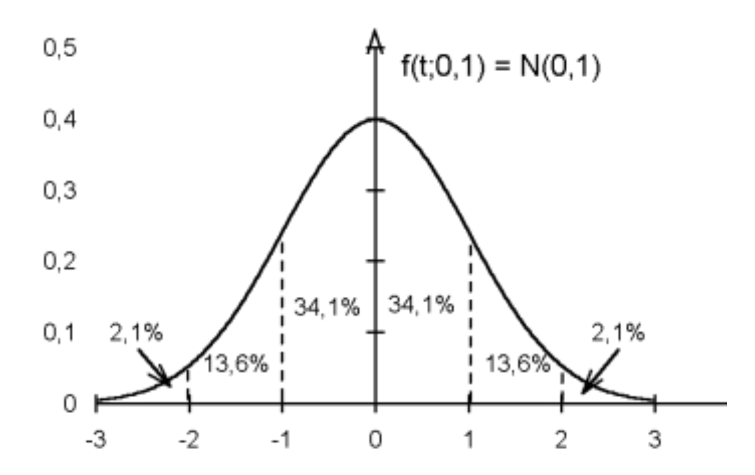


Figure 4 : loi normale centrée réduite N(0, 1)

Les probabilités correspondant aux divers intervalles ont été calculées et regroupées dans une table numérique. Ainsi la table A.1 (en fin de polycopié) permet, à partir d'une probabilité α donnée, de trouver les bornes $-u_\alpha$, $+u_\alpha$ d'un intervalle symétrique autour de 0, tel que

$$Pr(t \notin [-z_\alpha ; +z_\alpha]) = \alpha$$

ou encore, à partir de u_α, de trouver α.

D'où par exemple :

$$Pr(t \in [-z_\alpha ; +z_\alpha]) = 1 - \alpha$$

$$Pr(t > z_\alpha) = Pr(t < -z_\alpha) = \alpha/2$$

On observe ainsi que environ 68 % de la surface est comprise entre (-1 et +1), 95 % entre (-2 et +2) et 99 % entre (-3 et +3) (la table A.1 ne permet pas de trouver des valeurs aussi précises que celles de la figure 4).

Transformation d'une loi normale quelconque en loi normale centrée réduite

Soit une variable *X* distribuée selon une loi normale d'espérance μ et d'écart-type σ.

Alors la variable $t = \frac{X - \mu}{\sigma}$ est distribuée selon une loi normale centrée réduite.

Les probabilités obtenues pour la loi centrée réduite permettent de calculer les probabilités pour une loi normale quelconque, à l'aide de cette transformation :

$$t = \frac{X - \mu}{\sigma}.$$

Soit par exemple à calculer $Pr(a \le X \le b)$. Par la transformation, on a $Pr(a \le X \le b) = Pr(c \le t \le d)$ avec

$c = \frac{a - \mu}{\sigma}$ et $d = \frac{b - \mu}{\sigma}$.

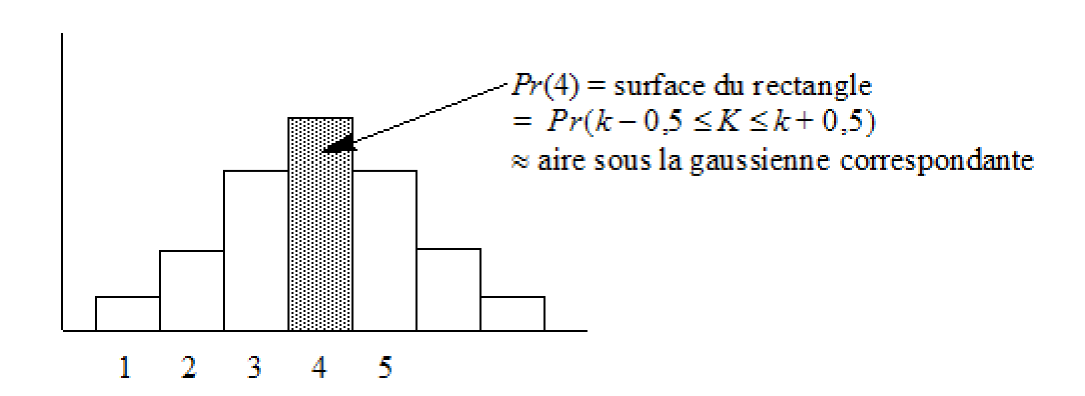
La probabilité cherchée, sur la variable *X*, revient donc à lire sur la table de la loi centrée réduite (variable *t*), la probabilité de se trouver entre *c* et *d*.
On remarque en particulier que $Pr(-2 \leq t \leq 2) = Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0{,}95$

Approximation de la distribution binomiale par la loi normale

Lorsque *n* est grand, et que Π et 1−Π ne sont pas trop proches de 0 (en pratique si $n\Pi \geq 5$ et $n(1-\Pi) \geq 5$), alors on constate que la distribution binomiale tend vers la distribution normale de moyenne *n*Π et de variance*n*Π(1−Π) ; plus précisément, pour une variable *K* distribuée selon une loi binomiale B(*n*, Π) et une variable *X*distribuée selon une loi normale N($\mu = n\Pi$, $\sigma^2 = n\Pi(1-\Pi)$), on a :

$$Pr(K = k) = Pr(k) = Pr(k - 0{,}5 \leq K \leq k + 0{,}5) \approx Pr(k - 0{,}5 \leq X \leq k + 0{,}5)$$

On choisit l'artifice de représenter graphiquement *Pr*(*k*) par un rectangle dont la base est [*k* - 0,5, *k* + 0,5] et la surface est *Pr*(*k*) pour comparer la loi discrète *Pr*(*k*) et la loi normale continue.



Approximation de la loi de Poisson par la loi normale

Lorsque son paramètre λ est grand (en pratique supérieur à 25), une loi de Poisson peut être approchée par une loi normale d'espérance λ et de variance λ.
Le principe est analogue à celui utilisé pour l'approximation de la loi binomiale par la loi normale.

7.2 - Lois continues

7.2.2 - Loi du χ^2 (chi-2)

7.2.2.1 Définition

C'est une loi dérivée de la loi normale, très importante pour ses applications en statistiques comme nous le reverrons dans les tests.

Soient $X_1, ..., X_n$ des variables aléatoires indépendantes, chacune étant distribuée selon une loi normale centrée réduite :

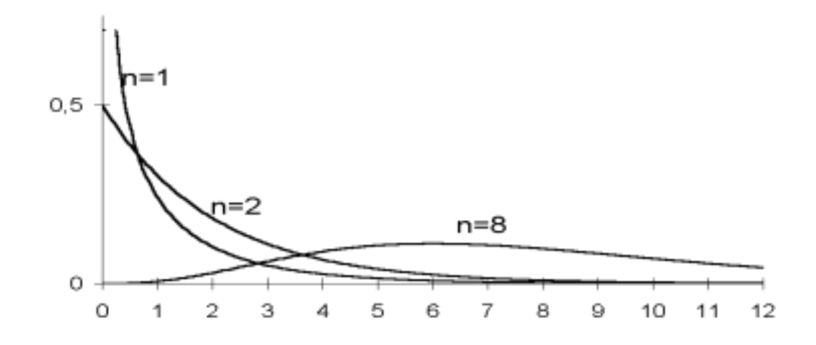
$\forall i, X_i \sim N(0, 1)$

La distribution de $S = X_1^2 + X_2^2 + ... + X_n^2$ (somme des carrés des X_i) est appelée loi de χ^2 à n degrés de liberté (en abrégé d. d. l.), que l'on note $\chi^2(n)$ où n est le nombre de d. d. l., seul paramètre de la loi.

Loi du $\chi^2(n)$	
Espérance	n
Variance	$2n$
Ecart-type	$\sqrt{2n}$

7.2.2.2 Propriétés

a. Allure de la distribution de $\chi^2(n)$ pour différentes valeurs de n



b. Pour n = 1, la courbe décroît de +∞ vers zéro de façon monotone ; pour n = 2, la courbe décroît de façon monotone de 0,5 à zéro ; pour $n > 2$, la courbe part de 0, a son maximum pour $x = n - 2$, puis redescend vers zéro.

c. Propriété asymptotique

La loi d'une variable X suivant un $\chi^2(n)$ tend vers une loi normale lorsque $n \to +\infty$. On a donc, après avoir centré et réduit cette variable :

$$\frac{X - n}{\sqrt{2n}} \sim N(0,\ 1)$$

NB : Dans la pratique, on utilise plutôt la variable $Y = \sqrt{2X} - \sqrt{2n-1}$ dont on montre qu'elle est à peu près distribuée selon une loi normale centrée réduite dès que $n > 30$.

d. Tables

De même que pour la loi normale centrée réduite, une table existe pour la loi du χ^2 (voir en fin de polycopié, table A.4). Cette table indique pour une probabilité α donnée, et un degré de liberté n donné, la valeur $K_{n,\alpha}$ telle que $Pr(X > K_{n,\alpha}) = \alpha$.

7.2 - Lois continues

7.2.3 Loi de Student (hors programme)

Il s'agit encore d'une loi dérivée de la loi normale, très utilisée dans les tests statistiques. On considère une première variable aléatoire X, distribuée selon une loi normale centrée réduite, puis une seconde variable Y, indépendante de X, distribuée selon un χ^2 à n degrés de liberté.

Alors la variable aléatoire $Z = \sqrt{n}\frac{X}{\sqrt{Y}}$ est distribuée selon une loi de Student à n degrés de liberté, notée t(n).

Loi de Student t(*n*)	
Espérance	0
Variance	$\frac{n}{n-2}$
Ecart-type	$\sqrt{\frac{n}{n-2}}$

La courbe correspondante est symétrique autour de 0, et son allure est proche de celle de la loi normale.

Cette loi est centrée, mais non réduite : la variance, $\frac{n}{n-2}$, est supérieure à 1.

Lorsque *n* croît, en pratique pour *n* > 30, la variance peut être prise égale à 1, et la distribution assimilée à celle d'une loi normale centrée réduite.

7.2.4 Loi exponentielle (hors programme)

Cette loi décrit par exemple le processus de mortalité dans le cas où le « risque instantané » de décès est constant. La loi correspondante est :

$f(x) = \lambda e^{-\lambda x}$ avec $\lambda > 0$ et $x \geq 0$

où *x* est la durée de vie.

Loi exponentielle	
Espérance	$1/\lambda$
Variance	$1/\lambda^2$
Ecart-type	$1/\lambda$

7.3 - Application de la Loi de Poisson à l'interprétation d'un risque sanitaire possible qui n'a pas encore été observé

7.3.1 Introduction

Dans de nombreux cas, on s'intéresse à un risque sanitaire a priori faible et on constate, après *n* observations, que l'événement redouté n'a jamais eu lieu. Par exemple, au bout de 10 000 prescriptions d'un médicament nouveau, on n'a pas observé un seul effet indésirable. Ou un chirurgien examinant le suivi de ses 50 dernières interventions avec une technique très innovante a la satisfaction de ne pas avoir eu un seul échec. Ces observations semblent plutôt rassurantes, mais que peut-on en tirer pour apprécier le risque encouru au bout de 1 000 000 de ces prescriptions, ou au bout de 100 000 interventions avec cette technique chirurgicale ?

C’est un problème qui se pose particulièrement en pharmacovigilance : on admet dans de nombreuses classes thérapeutiques qu’un médicament qui tuerait un malade sur 1 000 000 de prescriptions devrait être retiré du marché. Le biostatisticien posera donc le problème suivant : sachant qu’après 10 000 prescriptions aucun décès causé par le médicament n’a été observé, quel nombre de décès peut cependant être redouté sur 1 000 000 de prescriptions qui soit compatible avec ce qu’on sait aujourd’hui (0 décès sur 10 000). Si ce nombre dépasse 1 (on verra dans l’application numérique traitée plus loin qu’il le dépasse de loin), la plus grande vigilance s’imposera !

Le raisonnement qui suit combine le calcul des probabilités (loi de Poisson, approximation d’une loi binomiale par une loi de Poisson), et « l’inférence statistique ». L’inférence statistique est le mode de raisonnement qui permet à partir d’une observation (ici : 0 effet indésirable sur 10 000) de tirer des conclusions générales (ici : ce qui pourrait arriver sur 1 000 000 ou 10 000 000 de prescriptions). Les applications principales de l’inférence concernent l’estimation statistique et les tests statistiques ; elles sont traitées en détail dans les chapitres suivants du polycopié. Mais l’application détaillée ci-après initie bien à ce mode de raisonnement.

7.3.2 Le problème « direct »

Dans le problème direct on suppose que le risque d’effet indésirable chez un patient donné est connu, et on veut calculer la probabilité d’observer l’événement « 0 effet indésirable au bout des *n* premières observations ».

- Soit Π le risque individuel (= la probabilité) d’un effet indésirable après traitement.
- Soit *n* le nombre de patients traités.

Alors le nombre X de patients présentant un effet indésirable suit une loi binomiale de moyenne $\lambda = n\Pi$. Si Π est très petit, *n* très grand et $n\Pi$ assez petit (ces hypothèses seront toujours faites dans ces problèmes de risque sanitaire) la loi de X peut être approximée par une loi de Poisson de paramètre λ. En particulier, $Pr(X = 0) = e^{-\lambda}$ et $Pr(X = 1) = \lambda e^{-\lambda}$.

Remarque : on sait calculer exactement $Pr(X = 0) = (1- \Pi)^n$ en exprimant que chacune des n prescriptions doit être sans effet indésirable (probabilité 1- Π). Quand Π est très petit, et nΠ pas trop grand on montre que cette valeur est très proche de $e^{-n\Pi}$.

7.3.3 Problème inverse

C’est le problème qui se pose en pratique : On observe l’événement « x = zéro effet adverse sur *n* patients traités ». Que peut-on alors dire de Π ? (ou de λ car $\Pi = \lambda/n$). On approchera la distribution de X par la loi de Poisson de paramètre $\lambda = n\Pi$.

<u>Pour répondre, la méthode utilisée est d’écarter les valeurs de Π pour lesquelles notre observation (x = 0 après *n*répétitions) serait « invraisemblable ».</u>

Mais que veut dire invraisemblable ? Pour fixer les idées, imaginons que l’on ait observé x=0 parmi *n* = 10 000 traitements.

- Si Π avait été 1/1000, la probabilité de l'événement observé (« x = 0 sur 10 000 ») aurait été e^{-10} = 0,000045 = 4,5/100 000. (car $\lambda = n\Pi = 10$). Cette probabilité est infime : tout le monde sera d'accord pour dire qu'il aurait été très invraisemblable de n'avoir aucun effet indésirable si le risque individuel inconnu était de Π = 1/1000.
- Si Π avait été 1/10 000, la probabilité de l'événement observé aurait été e^{-1} = 0,37 = 37% (car $\lambda = n\Pi = 1$). La probabilité 0,37 est « forte ». Tout le monde sera d'accord pour dire qu'il n'était pas invraisemblable de n'observer aucun effet indésirable si le risque individuel était Π = 1/10 000.

e^{-10} est une « petite » probabilité - l'événement est invraisemblable ; e^{-1} est une « grande » probabilité - l'événement n'est pas invraisemblable. Pour préciser quantitativement ce que veut dire « invraisemblable » il faut choisir un seuil en dessous duquel on va déclarer qu'une probabilité est « petite » : la valeur conventionnelle retenue universellement est 5%. La traduction de ce choix est qu'on décide de ne pas trouver « invraisemblable » un événement ayant 10 chances sur 100 de se produire (car 10% > 5%), mais de trouver « invraisemblable » un événement ayant 1 chance sur 100 de se produire.

Ce choix permet d'apporter une solution au problème posé : l'observation « x = 0 sur 10000 » est invraisemblable si sa probabilité est inférieure à 5%, c'est-à-dire si $\exp(-n\Pi) < 5\%$. Toutes les valeurs de Π supérieures à la solution de cette inégalité seront réputées « invraisemblables » ; et les autres seront réputées « vraisemblables ».
La résolution de l'inéquation donne $\Pi > \Pi_{lim} = \ln(0,05)/n = 3/n$.

Résultat
Quand on observe 0 effet indésirable parmi *n* répétitions, ceci est compatible avec un risque individuel compris entre 0 et 3/*n*. En revanche, les risques supérieurs à 3/*n* sont jugés invraisemblables.

Comme la médecine veille à limiter le plus possible le risque encouru par les malades à l'occasion d'un traitement, la démarche de « précaution » est, lorsqu'on observe 0 effets indésirables sur *n* traitements, de conclure que le risque réel du traitement peut aller jusqu'à 3/*n*. Il peut bien sûr être plus petit : l'avenir le dira, et l'estimation du risque se précisera au fur et à mesure que de plus en plus de patients auront été traités. Mais en attendant, pour se préparer « au pire », on doit considérer la valeur maximum non invraisemblable, à savoir 3/*n*.

7.3.4 Application numérique

On traite 10 000 patients sans observer d'événement indésirable. Evaluer la limite supérieure du nombre d'événements indésirables qu'on peut redouter sur 1 000 000 de prescriptions, compatibles avec cette observation préliminaire.

Par application du résultat, l'observation « x = 0 sur 10,000 » est compatible avec un risque individuel maximum de 3/10000.

Sur 1 000 000 de prescriptions, on aura donc au maximum un nombre de
3/10 000 × 1 000 000 = 300 effets indésirables.

Ainsi, un laboratoire pharmaceutique ayant mis sur le marché une nouvelle molécule ne peut en rien être « rassuré » sur le risque associé à celle-ci après avoir constaté qu'il n'y avait aucun décès sur les 10 000 premières prescriptions.
Cette observation reste compatible avec un risque de 3/10000, et 300 décès sur le 1 000 000 de prescriptions suivant, ce qui serait une catastrophe sanitaire.

Ceci illustre également la difficulté de garantir un médicament « sans risque » lorsqu'il n'a été testé que chez quelques milliers de patients afin de voir s'il est efficace (voir chapitre 15 sur la méthodologie des études épidémiologiques et les essais thérapeutiques), comme c'est généralement le cas au moment de la demande d'autorisation de mise sur le marché.

7.3.5 Remarque

Le calcul menant au résultat encadré repose sur un raisonnement subtil (qui sera retrouvé dans les chapitres relatifs aux tests statistiques).

Chacun doit comprendre dans l'exemple traité que le résultat trouvé ne signifie pas qu'il y a 5 chances sur 100 pour que le risque sanitaire soit de 3/10 000 (cette interprétation fausse du « 3 » est très majoritairement faite, y compris par de nombreux professionnels). Ce que le calcul indique, c'est que si le risque de mort était de 3/10000 (il ne l'est peut-être pas), il y aurait 5 chances sur 100 d'observer 0 décès sur 10 000 prescriptions, comme on l'a fait. En reprenant les notations des probabilités conditionnelles, il ne faut pas confondre Pr(A / B) et Pr(B / A), avec ici $A = \{\lambda = 3\}$ et $B = \{X = 0\}$.

Chapitre 8 - Statistiques descriptives

Les statistiques descriptives visent à représenter des données dont on veut connaître les principales caractéristiques quantifiant leur variabilité.

8.1 Rappels et compléments

On suppose que l'on s'intéresse à une caractéristique particulière observable chez des individus issus d'une population ; cette caractéristique sera appelée variable ; si cette caractéristique peut varier entre les individus, sans pouvoir l'anticiper, on l'appellera **variable aléatoire**. Le dispositif permettant d'obtenir une valeur de la variable est l'**expérience aléatoire**. Cette définition imagée est compatible avec la définition du chapitre 6.

Rappel

Il existe deux grands groupes de variables :

a. Les variables **quantitatives** qui sont des variables ordonnées, productives de nombres. Exemples : nombre d'enfants dans une famille, glycémie, taille d'un individu, nombre de colonies bactériennes dans un milieu de culture.
   Parmi ces variables quantitatives, certaines prennent un continuum de valeurs (entre deux valeurs possibles, il existe toujours une troisième valeur possible) ; ces variables sont dites **continues**. D'autres ne prennent que des valeurs discontinues ; elles sont dites **discrètes**, finies ou non.
b. Les variables **qualitatives** qui produisent des valeurs non numériques. Exemples : sexe, couleur des cheveux, appartenance au groupe des fumeurs ou des non fumeurs, présence ou absence d'une maladie.
   Les valeurs peuvent être ordonnées ; on parle alors de variable qualitative **ordinale**. Exemple : intensité d'une douleur (faible, moyenne, forte).
   Si les valeurs ne peuvent pas être ordonnées, il s'agit d'une variable **catégorielle** (ou **nominale**).

Remarque

L'individu évoqué ci-dessus, sur lequel on observe les caractéristiques d'intérêt, la variable, n'est pas nécessairement un individu physique. C'est l'entité sur laquelle s'opère l'observation de la variable d'intérêt. Exemples : famille, colonies bactériennes.

Définition

L'entité sur laquelle peut s'observer la variable aléatoire s'appelle l'**unité statistique**. Connaître le phénomène mettant en jeu cette variable, ou connaître cette variable, c'est connaître la probabilité pour qu'un individu tiré au hasard dans la population présente telle valeur de la variable. On peut apprécier la probabilité d'un événement aléatoire grâce à l'interprétation suivante de la notion de probabilité. Cette interprétation est cohérente avec les cours précédents.
On **interprétera** la probabilité d'un événement aléatoire comme la valeur limite de la fréquence avec laquelle l'événement se réalise au cours d'un nombre **croissant** de

répétitions de l'expérience. Autrement dit comme la valeur limite du rapport du nombre de fois où l'événement s'est réalisé et du nombre de répétitions de l'expérience.

Remarques

- Ce qui précède peut être vu comme une interprétation de la notion de probabilité (voire comme une définition).
- En dépit de cette interprétation, la probabilité d'un événement aléatoire reste
  - une fiction
  - du domaine théorique.

  Mais cette interprétation a deux conséquences :

  - pour approcher une probabilité on est amené à répéter une expérience,
  - les fréquences se substituent aux probabilités ; elles seront les contreparties des probabilités.

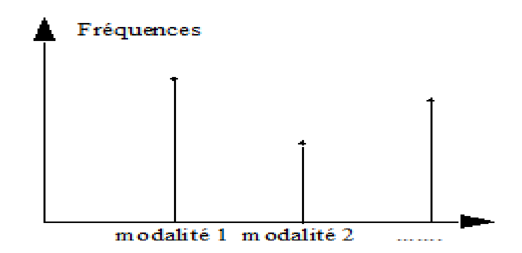
  On va donc répéter une expérience un nombre fini de fois, noté n ; on aura donc observé une sous-population appelée **échantillon**. Chaque expérience aléatoire produit un résultat x_i ; on disposera donc de x_1, ..., x_n, ensemble appelé **échantillon de valeurs** de la variable étudiée X.

- De façon plus formelle, on définit un **échantillon d'une variable aléatoire** de la manière suivante :
  Un **échantillon** de taille n **d'une variable aléatoire** X est un ensemble X_1, X_2, ..., X_n de n variables aléatoires, indépendantes entre elles, et ayant chacune la même distribution que X. On peut donc dire qu'un échantillon de valeurs de X est **une** réalisation de l'échantillon de la variable X tel qu'il vient d'être défini.

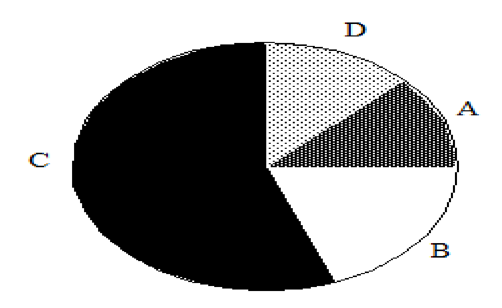
8.2 - Représentation complète d'une série d'expériences

8.2.1 Cas d'une variable qualitative

La variable est décrite par la suite des probabilités des différentes modalités. Si l'on connaissait ces probabilités, on produirait le diagramme en bâtons (ou répartition « vraie ») de cette variable ; on va produire la **répartition observée** par substitution aux probabilités inconnues des fréquences observées. Si la variable est ordinale, on respectera cet ordre dans l'énumération des modalités portées en abscisses.

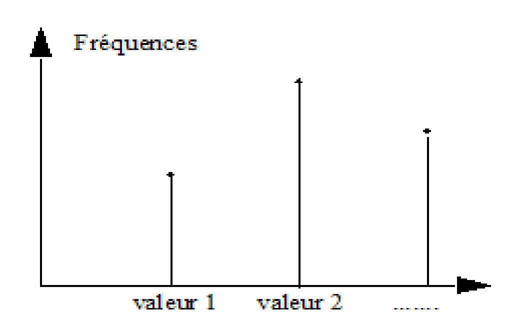


D'autres types de représentation sont utilisés : par exemple la représentation en camembert où les différentes modalités sont représentées par secteurs angulaires d'angles au centre proportionnels aux fréquences observées.



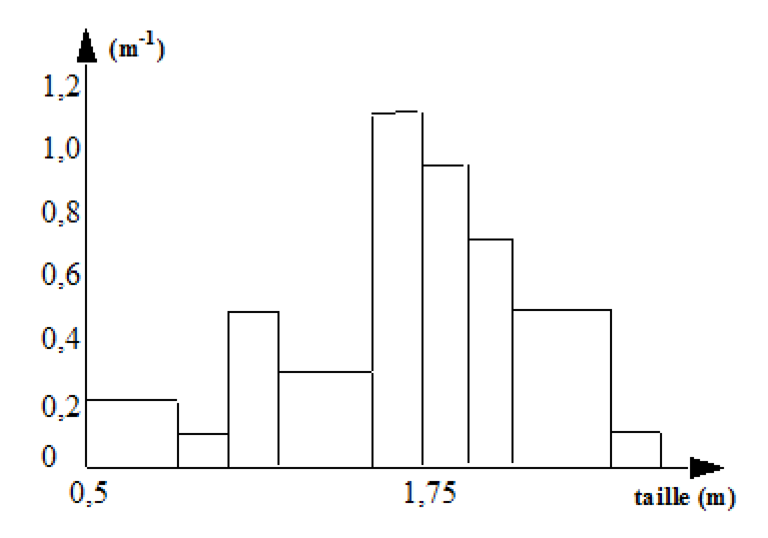
8.2.2 Cas d'une variable quantitative discrète

La situation est similaire si ce n'est qu'il existe un ordre et une échelle naturels en abscisses ; la répartition observée se nomme également histogramme en bâtons.



8.2.3 Cas d'une variable quantitative continue. Notion d'HISTOGRAMME

Dans le cas de variables continues, on va choisir de représenter les données graphiquement d'une façon qui soit proche de la représentation d'une densité de probabilité d'une variable aléatoire continue. Pour cela on découpe l'ensemble du domaine des valeurs possibles de la variable étudiée en intervalles contigus dont on choisit le nombre et les bornes. Afin d'obtenir une représentation proche d'une densité de probabilité, on décide de représenter indirectement la fréquence des valeurs observées comprises entre deux bornes consécutives par la surface d'un rectangle dont la base sera précisément cet intervalle. Autrement dit la hauteur de ce rectangle sera le rapport de la fréquence observée de ces valeurs et de la différence entre ces bornes (différence également appelée largeur de la classe).



Les bornes sont choisies arbitrairement ; néanmoins, pour que l'histogramme ait un sens il est nécessaire que la taille de chaque classe constituant un intervalle comprenne un nombre suffisamment grand de valeurs observées, de telle façon que la surface d'un rectangle élémentaire puisse être interprétée comme approchant la probabilité pour que la variable prenne une valeur comprise dans l'intervalle du rectangle. Si la taille de l'échantillon croît, la surface de chaque rectangle tend vers la probabilité que la variable ait une valeur incluse dans l'intervalle correspondant. De plus, si la taille *n* de l'échantillon est grande, on peut alors sans inconvénient construire un plus grand nombre de classes, c'est-à-dire construire par exemple deux fois plus de rectangles, chacun ayant un support deux fois plus petit. En répétant cette opération, *n* croissant, on peut comprendre que l'histogramme tend (d'une façon que nous ne préciserons pas ici) vers la densité de probabilité de la loi qui a généré l'échantillon.

8.3 - Représentation simplifiée d'une série d'expériences

On a défini certains indicateurs pour représenter, de façon plus résumée que ci-dessus, un échantillon de valeurs issues d'une variable aléatoire.

Les indicateurs présentés ci-dessous ne concernent que les variables quantitatives.

8.3.1 Indicateurs de localisation des valeurs

Médiane observée

C'est la valeur qui partage l'échantillon en deux groupes de même effectif ; pour la calculer, il faut commencer par ordonner les valeurs (les ranger par ordre croissant par exemple)

Exemple : soit la série 12 3 24 1 5 8 7

on l'ordonne : 1 3 5 **7** 8 12 24

7 est la médiane de la série

Moyenne observée

C'est l'indicateur de localisation le plus fréquemment utilisé. La moyenne observée d'un échantillon de *n* valeurs x_1, ..., x_n est définie comme la moyenne arithmétique de ces valeurs ; on la note souvent m_x, ou simplement *m* s'il n'y a pas de confusion possible :

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Avec la série précédente, qui comporte *n* = 7 valeurs, on obtient :

$$m = \frac{1}{7}\sum_{i=1}^{7} x_i = \frac{12+3+24+1+5+8+7}{7} = 8,57$$

8.3.2 Indicateurs de dispersion des valeurs

Variance observée

La variance observée d'un échantillon $\{x_i\}$ $i = 1, ..., n$ est donnée par

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - m)^2$$

Attention : on divise par *n* -1 et non par *n* pour que la variance observée soit un bon estimateur de la variance théorique de la loi (nous reverrons ce point dans la suite). Une autre expression de s^2, équivalente, est indiquée dans le résumé de ce chapitre.

Ecart-type observé

L'écart-type observé, noté *s*, est défini par $s = \sqrt{s^2}$.

8.4 - Reformulation de la moyenne et de la variance observées

Prenons le cas d'une variable quantitative discrète.

Les données sont notées x_1, ..., x_n.

Les *k* valeurs possibles de la variable sont notées $val_1, val_2, ..., val_k$.

Exemple d'un jet de dé : $val_1 = 1, ..., val_6 = 6$

Chaque donnée x_i coïncide avec une certaine valeur val_j

Par exemple pour le jet de dé, on peut avoir

- jet n?1 ; $x_1 = 1 = \text{val}_1$
- jet n?2 ; $x_2 = 1 = \text{val}_1$
- jet n?3 ; $x_3 = 4 = \text{val}_4$
- jet n?4 ; $x_4 = 3 = \text{val}_3$
- jet n?5 ; $x_5 = 6 = \text{val}_6$
- jet n?6 ; $x_6 = 1 = \text{val}_1$
- jet n?7 ; $x_7 = 2 = \text{val}_2$
- jet n?8 ; $x_8 = 5 = \text{val}_5$
- jet n?9 ; $x_9 = 6 = \text{val}_6$

Alors : $\sum_{i=1}^{n} x_i = \sum_{j=1}^{k} n_j \text{val}_j$

où n_j est le nombre de fois où une observation coïncide avec val_j

Dans notre exemple du jet de dé, on a : $n_1 = 3$, $n_2 = 1$, $n_3 = 1$, $n_4 = 1$, $n_5 = 1$, $n_6 = 2$

Finalement $m = \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{j=1}^{k} \frac{n_j}{n}\text{val}_j$

Mais $\frac{n_j}{n}$ est une approximation de $Pr(\text{face marquée} = \text{val}_j)$

Ainsi m est une estimation - une appréciation - de :

$$\sum_j \text{val}_j Pr(\text{valeur de la variable} = \text{val}_j)$$

c'est-à-dire une appréciation de l'espérance mathématique de la variable.

On raccorde ainsi une moyenne observée à une grandeur descriptive du phénomène étudié, à une grandeur dite « théorique » ou « vraie ».

On peut dire ceci : la répétition des expériences vise à estimer Pr(valeur de la variable = certain niveau). La moyenne observée permet d'estimer quelque chose de plus grossier, une combinaison de toutes ces probabilités, précisément l'espérance mathématique

$$\mu = \sum_j \text{val}_j Pr(\text{valeur de la variable} = \text{val}_j)$$

C'est la raison pour laquelle dans la suite on utilisera également la terminologie **MOYENNE « VRAIE »** ou**MOYENNE THEORIQUE** de la variable pour parler de l'espérance mathématique.

Retenons :

ESPERANCE MATHEMATIQUE,

MOYENNE « VRAIE »,

MOYENNE THEORIQUE

sont SYNONYMES. Ce sont des grandeurs théoriques.

Remarque

La même analyse peut être faite - mais l'expression est un peu plus délicate - dans le cas d'une variable quantitative continue. La moyenne observée approxime là encore l'espérance mathématique.

8.4.2 Reformulation de la variance observée

De la même façon on peut obtenir le résultat suivant : s^2 est une approximation de la grandeur $\sigma^2 = \sum_j (\text{val}_j - \mu)^2 Pr(\text{valeur de la variable} = \text{val}_j)$

Cette expression, introduite dans le chapitre 6 sous le nom de variance sera souvent dénommée dans la suite**VARIANCE « VRAIE »** ou **VARIANCE THEORIQUE** de la variable.

Dans le cas d'une variable continue, la variance observée s^2 approxime :

$$\sigma^2 = \int_{\Re} (x - \mu)^2 f(x)\,dx$$

LES DIFFERENCES ENTRE CES NOTIONS DE MOYENNE ET VARIANCE « VRAIES », ET DE MOYENNE ET VARIANCE OBSERVEES SONT **ESSENTIELLES** ; NOUS ENGAGEONS LE LECTEUR A BIEN LES COMPRENDRE AVANT DE POURSUIVRE.

8.5 - Cas particulier d'une variable à deux modalités - Proportion

On est très souvent amené à considérer des variables à deux modalités, c'est-à-dire des expériences aléatoires à deux événements élémentaires.

Exemples :

- maladie : maladie présente - maladie absente
- signe clinique : présent - absent
- traitement : individu traité - individu non traité

Or on peut transformer une telle variable en variable quantitative, sans restriction de généralité, par un artifice de codage :

- une des modalités est codée avec la valeur numérique 0 ;
- l'autre modalité est codée avec la valeur numérique 1.

Une telle variable s'appelle variable de **Bernoulli**.

Notons *X* cette variable.

Elle est complètement décrite par la donnée de *Pr*(valeur de la variable = 1) car

Pr(valeur de la variable = 1) + *Pr*(valeur de la variable = 0) = 1.

On utilise la notation conventionnelle suivante : *Pr*(valeur de la variable = 1) SE NOTE Π.

8.5.1 Expression de l'espérance mathématique de *X*

Utilisant l'expression générale de l'espérance mathématique, et remarquant que $val_1 = 0$, $val_2 = 1$, on obtient :

$$\mu = \sum_j val_j Pr(\text{valeur de la variable} = val_j) = 0 \times (1-\Pi) + 1 \times \Pi = \Pi$$

Ainsi, μ = Π = *Pr*(valeur de la variable = 1) = probabilité de la modalité codée 1 = PROPORTION VRAIE des individus présentant la modalité 1.

8.5.2 Expression de la variance de *X*

$$\sigma^2 = \sum_j (val_j - \mu)^2 Pr(\text{valeur de la v.a.} = val_j) = (0-\Pi)^2(1-\Pi) + (1-\Pi)^2\Pi = \Pi(1-\Pi)$$

8.5.3 Interprétation de la moyenne observée

$$m = \frac{1}{n}\sum_i x_i = \frac{1}{n}[0+0+1+0+1+1+\ldots] = \frac{\text{nombre de fois où } X = 1}{n}$$

Ainsi, *m* coïncide avec la fréquence observée de la modalité codée 1. Cette fréquence sera notée *p* et s'appelle de façon naturelle PROPORTION OBSERVEE d'individus présentant la modalité 1.

Exemple

Dans le cas de l'étude d'un signe clinique, en codant 1 la présence du signe clinique, *m* (donc *p*) sera la fréquence observée de la présence du signe ou encore le pourcentage des individus présentant le signe (à un facteur 100 près).

En résumé

- si *X* est une variable de Bernoulli,
    - sa moyenne « vraie » = Π
    - sa variance « vraie » = Π(1 - Π)
- UNE PROPORTION OBSERVEE EST UNE MOYENNE OBSERVEE.

8.6 - Conclusion : la variable aléatoire moyenne arithmétique

On a jusqu'ici associé une valeur de moyenne observée à une série de *n* réalisations d'une variable aléatoire quantitative *X*. Mais chaque expérience consistant à recueillir *n* réalisations de la variable *X* permet de calculer une valeur, différente à chaque expérience, de moyenne observée. Autrement dit, la moyenne observée doit être vue comme une nouvelle variable aléatoire que nous appellerons moyenne arithmétique ; on la notera *M*. Dans certains cas, afin de rappeler que cette variable dépend de *n*, on notera M_n la variable construite à partir de *n* réalisations de *X*.

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

On utilisera la terminologie suivante :

on dira que *M* (ou M_n si nécessaire) est la VARIABLE ALEATOIRE MOYENNE ARITHMETIQUE DEDUITE DE LA VARIABLE ALEATOIRE *X*, FONDEE SUR *n* REPETITIONS

ou, de façon équivalente que *M* (ou M_n si nécessaire) est la VARIABLE ALEATOIRE MOYENNE ARITHMETIQUE ASSOCIEE A LA VARIABLE ALEATOIRE *X*, FONDEE SUR *n* REPETITIONS

Remarque

Dans le cas où *X* est une variable de Bernoulli, M_n sera notée P_n (et *M* simplement *P*). Il s'agit d'une variable aléatoire proportion dont on connaît déjà pratiquement la distribution puisque

$nP_n \sim B(n, \Pi)$ (voir section 7.1.2).

Chapitre 9 - Fluctuations de la moyenne observée : la variable aléatoire moyenne arithmétique

On conserve le contexte d'étude du chapitre précédent, c'est-à-dire l'examen de la variabilité d'une grandeur (variable aléatoire) dans une population d'individus ou unités statistiques. Mais on s'intéresse ici à la variable aléatoire « moyenne arithmétique ».

9.1 Première propriété de la variable aléatoire moyenne arithmétique

9.1.1 Un exemple

Prenons à nouveau le cas d'une variable discrète pouvant prendre les deux valeurs 0 et 1 [c'est-à-dire variable associée à présence-absence ou oui-non]. Supposons que l'on ait des raisons de penser que $Pr(X = 0) = Pr(X = 1) = 1/2$. On a vu qu'une telle variable a pour espérance 1/2, pour variance « vraie » 1/4.

On peut, par le calcul, pronostiquer le résultat d'une répétition d'expériences. En particulier, calculer la répartition de la variable « moyenne arithmétique calculée sur un échantillon de deux individus », notée M_2, ici deux lancers de pièce.

On isole cette variable. Quelles valeurs peut-elle prendre, avec quelles probabilités ?

jet 1 : résultats	Proba jet 1	jet 2 : résultats	Proba jet 2	Proba jet1, jet2	M_2
0	1/2	0	1/2	1/4	1/2(0+0) = 0
0	1/2	1	1/2	1/4	1/2(0+1) = 1/2
1	1/2	0	1/2	1/4	1/2(1+0) = 1/2
1	1/2	1	1/2	1/4	1/2(1+1) = 1

Ainsi, $Pr(M_2 = 0) = \frac{1}{4}, Pr\left(M_2 = \frac{1}{2}\right) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}, Pr(M_2 = 1) = \frac{1}{4}$

Alors :

- $\text{moyenne vraie de } M_2 = 0 \times \frac{1}{4} + \frac{1}{2} \times \frac{1}{2} + 1 \times \frac{1}{4} = \frac{1}{2} = \text{moyenne vraie de } X$

- $\text{variance vraie de } M_2 = \left(0-\frac{1}{2}\right)^2 \times \frac{1}{4} + \left(\frac{1}{2}-\frac{1}{2}\right)^2 \times \frac{1}{2} + \left(1-\frac{1}{2}\right)^2 \times \frac{1}{4} = \frac{1}{8} = \frac{1}{2} \times \frac{1}{4}$

Ainsi la variance « vraie » de la moyenne arithmétique est plus faible que la variance « vraie » de la variable d'origine (la moitié ici). L'espérance reste inchangée. Et ainsi vont les choses si la taille des échantillons (ici 2) qui constituent les unités statistiques augmente. La dispersion de M diminue au fur et à mesure que M se trouve calculée sur la base d'un échantillon de taille croissante. Le « comment » de cette situation peut être résumé ainsi : les valeurs de la moyenne arithmétique deviennent de plus en plus probables dans un voisinage de l'espérance car le nombre de situations pouvant donner une valeur observée proche de l'espérance augmente dans ce voisinage. Cela est dû au fait que l'espérance mathématique est « au milieu » des valeurs possibles. On le voit sur l'exemple ci-dessus où l'espérance est obtenue dans les deux cas (0, 1) et (1, 0). C'est encore plus perceptible sur l'exemple d'un dé. Pour que la moyenne observée calculée sur deux jets de dé soit 6, il faut obtenir le résultat (6, 6) ; pour qu'elle soit 3, il faut un total de 6, c'est-à-dire (5, 1), (4, 2), (3, 3), (2, 4), (1, 5), soit un événement 5 fois plus probable.

Il est possible de quantifier tout cela. On peut généraliser ce qui a été obtenu avec deux jets de pièces et on obtient, quelle que soit la distribution de la variable étudiée - qu'elle soit continue ou discrète - les résultats fondamentaux suivants.

9.1.2 Généralisation

i. L'espérance mathématique, ou moyenne « vraie », de la variable aléatoire moyenne arithmétique calculée sur un échantillon de taille *n* coïncide avec la moyenne « vraie » de la variable étudiée, ce que l'on peut résumer par :

$$E(M_n) = E(X)$$

ii. La variance « vraie » de la variable aléatoire moyenne arithmétique calculée sur un échantillon de taille *n* est égale à la variance « vraie » de la variable **DIVISEE PAR** *n*, ce que l'on peut résumer par :

$$\sigma^2(M_n) = \frac{1}{n}\sigma^2(X)$$

d'où la relation entre écarts-types :

$$\sigma(M_n) = \frac{1}{\sqrt{n}}\sigma(X)$$

iii. Dans le cas où *X* est une variable de Bernoulli de paramètre Π ($Pr(X = 1) = \Pi$), les relations précédentes deviennent :

$$\mu(P_n) = \Pi$$

$$\sigma^2(P_n) = \frac{\Pi(1-\Pi)}{n}$$

9.2 - Seconde propriété de la variable aléatoire moyenne arithmétique : le théorème central limite

On souhaiterait comparer, par curiosité, les distributions de plusieurs moyennes arithmétiques, correspondant à diverses variables aléatoires. Par exemple la taille, la glycémie. Ces distributions sont différentes, ne serait-ce qu'à cause des différences entre moyennes et variances « vraies ». Pour s'abstraire de ces premières différences, considérons la variable centrée réduite associée, soit pour chaque variable considérée :

$\frac{M_n - \mu(M_n)}{\sigma(M_n)}$ soit $\frac{M_n - \mu(X)}{\frac{\sigma(X)}{\sqrt{n}}}$ Maintenant toutes ces variables ont en commun leur espérance (0) et leur variance (1). Il se passe quelque chose d'extraordinaire : lorsque *n* est suffisamment grand, elles finissent par avoir en commun leur distribution, leur densité de probabilité.

Cela signifie que les distributions de toutes ces variables (moyennes arithmétiques centrées réduites issues de variables aléatoires différentes) finissent par coïncider, lorsque *n* est suffisamment grand, avec une distribution particulière unique. Cette distribution s'appelle **LOI NORMALE**, et puisque sa moyenne « vraie » est nulle et sa variance « vraie » est 1, on l'appelle LOI NORMALE CENTREE REDUITE ou encore distribution de Gauss ou de Laplace-Gauss (1800).

On la notera schématiquement N(0, 1) où 0 rappelle la valeur de la moyenne « vraie », 1 la valeur de la variance « vraie ».
Donc la propriété ci-dessus - connue sous le nom de théorème central limite - s'énonce :
THEOREME CENTRAL LIMITE
Soit *X* une variable aléatoire quantitative d'espérance mathématique μ, de variance « vraie » σ^2. Soit M_n la variable aléatoire moyenne arithmétique associée à *X* construite sur *n* répétitions.

La distribution limite de la variable aléatoire $\frac{M_n - \mu}{\frac{\sigma}{\sqrt{n}}}$ est la distribution normale centrée réduite notée N(0,1).

Il faut bien mesurer la portée de cette propriété. Quel que soit le phénomène étudié - apprécié par la variable aléatoire que l'on étudie - il suffit de connaître la moyenne et la variance de la variable pour déduire la **distribution** (la densité de probabilité) - c'est-à-dire l'expression la plus achevée des propriétés de variabilité - de la variable aléatoire moyenne arithmétique calculée sur un échantillon de taille suffisante. Nous reviendrons plus loin, au paragraphe résumé et précisions, sur cette notion vague « taille suffisante ». Or c'est peu de connaître moyenne, variance (ou écart-type) seulement - ex. : pour le poids à la naissance $\mu = 3$ kg, $\sigma = 1{,}2$ kg.

9.3 - Etude de la distribution normale (rappel)

La distribution limite que l'on a mise en évidence dépeint une variable aléatoire d'espérance mathématique 0 et de variance « vraie » 1, que l'on a appelée distribution normale centrée réduite ou N(0, 1).

La densité de probabilité est donnée par une fonction d'équation $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ et dont l'allure est représentée sur la figure 5.

Ses principales caractéristiques morphologiques sont les suivantes :

- elle est symétrique,
- elle présente deux points d'inflexion en $x = 1$ et $x = -1$

Par ailleurs, pour faciliter les calculs de probabilité relatifs à cette variable, des tables ont été construites qui donnent le lien entre α et u_α, où ces valeurs ont le sens suivant (voir figure 5) :

$$Pr(X \notin [-u_\alpha\,;+u_\alpha]) = \alpha$$

En particulier, pour $\alpha = 0{,}05$, la valeur u_α lue dans la table est 1,96, d'où $u_{0,05} = 1{,}96$

On peut voir facilement que toute probabilité $Pr(X \in [a,b])$ s'obtient à partir d'une telle table, quelles que soient les valeurs de a et b.

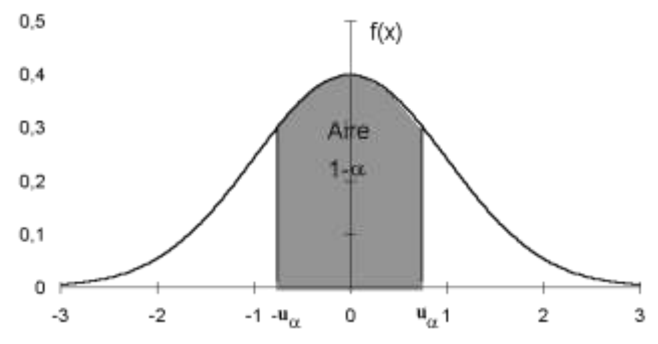


Figure 5 : loi normale centrée réduite

Remarque

Sur la base de cette loi centrée réduite, on définit toute une famille de lois de la façon suivante :

Si X est distribuée selon une loi normale centrée réduite (notation $X \sim N(0, 1)$), alors la variable $Y = \sigma X + \mu$, dont l'espérance est μ et la variance σ^2, est distribuée selon une loi normale d'espérance μ et de variance σ^2.
On écrit $Y \sim N(\mu, \sigma^2)$

A l'inverse, si on dit que $X \sim N(\mu, \sigma^2)$

cela veut dire que $\frac{X-\mu}{\sigma} \sim N(0, 1)$ (variable centrée réduite associée).

Exemple

La figure 6. présente l'aspect de deux distributions normales l'une N(0, 1), l'autre N(2,9 , 4).

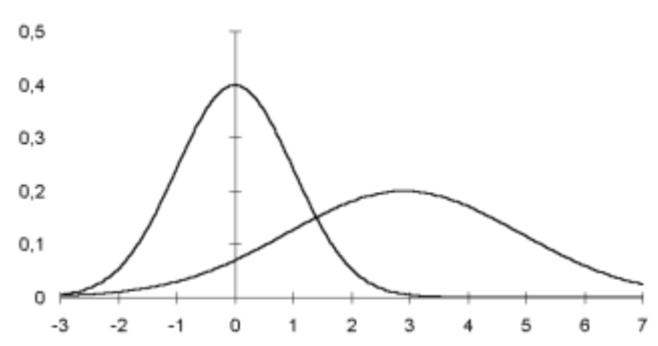


Figure 6 : exemple de lois normales

Résumé et précisions (théorème central limite)

Si n est suffisamment grand, X ayant pour moyenne « vraie » μ, pour variance « vraie » σ^2, alors :

$$\frac{M_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1) \text{ (à peu près)}$$

ou, de façon équivalente, $M_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ (à peu près)

où la notation ~ se lit : « est distribué comme » ou « suit une distribution ».

a. La distribution de M_n est exactement une loi normale (la mention *à peu près* est inutile), quel que soit *n*, si X elle-même est gaussienne (i.e. est distribuée normalement).
b. si X n'est pas gaussienne :
   - si X est une variable quantitative autre que Bernoulli, la condition de validité usuelle est $\mathbf{n \geq 30}$

- si X est une variable de Bernoulli (valeurs 0 et 1), la condition usuelle de validité est

$$\begin{cases} n\Pi \geq 5 \text{ et} \\ n(1-\Pi) \geq 5 \end{cases}$$

En outre dans ce cas, $\mu = \Pi$, $\sigma^2 = \Pi\,(1 - \Pi)$ si bien que l'on aura :

$$\frac{P_n - \Pi}{\sqrt{\frac{\Pi(1-\Pi)}{n}}} \sim \mathrm{N}(0,\,1) \text{ (à peu près)}$$

ou, de façon équivalente, $P_n \sim \mathrm{N}\left(\Pi,\, \frac{\Pi(1-\Pi)}{n}\right)$ (à peu près)

9.4 - Application du théorème central limite. Intervalle de Pari (I. P.)

9.4.1 - Définition de l'intervalle de pari (I. P.) d'une moyenne observée

On considère une variable aléatoire de moyenne « vraie » μ et de variance « vraie » σ^2.

On sait que pour n grand ($n \geq 30$, ou $n\Pi$ et $n(1 - \Pi) \geq 5$) :

la variable $Z = \frac{M_n - \mu}{\frac{\sigma}{\sqrt{n}}}$ est approximativement distribuée selon N (0, 1).

On se pose le problème suivant. On s'apprête à réaliser une série d'expériences, c'est-à-dire à mesurer la variable Xsur un échantillon de n individus. Peut-on construire un intervalle [a, b] tel que la probabilité pour que la moyenne observée que l'on s'apprête à calculer appartienne à cet intervalle ait une valeur donnée ? Il s'agit donc de construire un intervalle qui contienne avec une probabilité fixée la valeur observée que l'on va obtenir.

Il s'agit donc de trouver deux valeurs a et b telles que $Pr(a \leq M_n \leq b) = \text{valeur donnée} = 1 - \alpha$.

Exemple : $Pr(a \leq M_n \leq b) = 0,95$

Un tel intervalle [a, b] s'appelle **INTERVALLE DE PARI (I. P.)** de niveau 1 - α, ou encore intervalle de pari au risque α, ou encore **INTERVALLE DE FLUCTUATION**

La figure 7 illustre le problème posé.

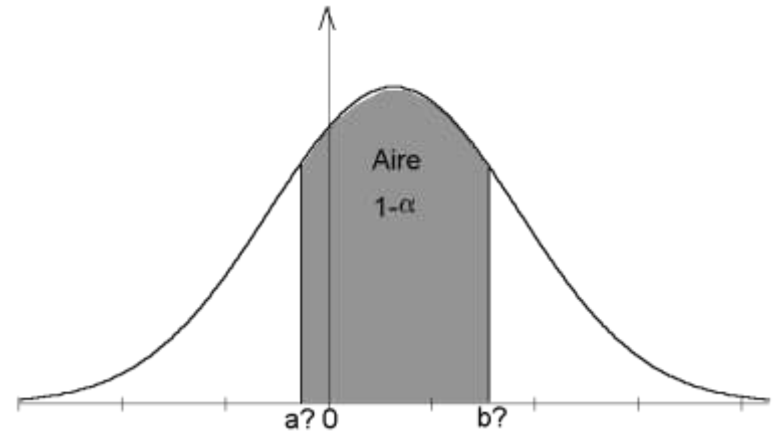


Figure 7 : le problème de l'intervalle de pari

Ce problème admet plusieurs solutions : sauf besoin spécifique on choisit un intervalle symétrique autour de μ (ce qui est naturel compte tenu de la distribution de M_n).

Résolution : $a = \mu - \lambda \frac{\sigma}{\sqrt{n}}$ et $b = \mu + \lambda \frac{\sigma}{\sqrt{n}}$

La valeur λ inconnue doit vérifier :

$$Pr\left(\mu - \lambda \frac{\sigma}{\sqrt{n}} \leq M_n \leq \mu + \lambda \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$Pr\left(-\lambda \frac{\sigma}{\sqrt{n}} \leq M_n - \mu \leq \lambda \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$Pr\left(-\lambda \leq \frac{M_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \lambda\right) = 1 - \alpha$$

Si le théorème central limite s'applique, l'expression ci-dessus suit une loi N(0, 1) ; notons-la Z. Alors λ doit vérifier $Pr(-\lambda \leq Z \leq \lambda) = 1 - \alpha$. C'est le u_α de la table.

Finalement : $\lambda = u_\alpha$

$$Pr\left(\mu - u_\alpha \frac{\sigma}{\sqrt{n}} \leq M_n \leq \mu + u_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$ et

$$\mathrm{IP}_{1-\alpha} = \left[\mu - u_\alpha \frac{\sigma}{\sqrt{n}} \; ; \; \mu + u_\alpha \frac{\sigma}{\sqrt{n}}\right]$$

Intervalle de Pari (I. P.) de la moyenne observée d'une variable de moyenne « vraie » μ, de

variance « vraie » σ^2 construite sur un échantillon de taille *n*

Exemple : α = 0,05 u_α= 1,96 $$IP_{0,95} = \left[\mu - 1,96\frac{\sigma}{\sqrt{n}} ; \mu + 1,96\frac{\sigma}{\sqrt{n}}\right]$$

Les conditions de validité de cette construction sont celles du théorème central limite, c'est-à-dire $n \geq 30$ pour les variables continues non normales et $n\Pi$, $n(1 - \Pi) \geq 5$ pour les variables de Bernoulli.

Cas d'une variable de Bernoulli : μ est notée Π, $\sigma^2 = \Pi(1 - \Pi)$. Donc

$$IP_{0,95} = \left[\Pi - 1,96\sqrt{\frac{\Pi(1-\Pi)}{n}} ; \Pi + 1,96\sqrt{\frac{\Pi(1-\Pi)}{n}}\right]$$

L'interprétation de l'intervalle de pari est fondamentale. Si cet intervalle est bien calculé, on est quasi sûr, avec une probabilité 1 - α (ici 0,95), d'obtenir une valeur de la moyenne observée comprise dans cet intervalle. En pariant que la valeur va tomber dans cet intervalle, on se trompera (en moyenne) dans cinq pour cent des expériences.

Exemple :

On a des raisons de penser que la fréquence d'une maladie dans la population est Π = 0,2. L'intervalle de pari de la moyenne observée (proportion observée) calculée sur 64 individus au niveau 0,95 est :

$$IP_{0,95} = \left[0,2 - \frac{1,96\sqrt{0,2(1-0,2)}}{\sqrt{64}} ; 0,2 + \frac{1,96\sqrt{0,2(1-0,2)}}{\sqrt{64}}\right] = [0,10 ; 0,30]$$

Il y a 95 chances sur 100 pour que la proportion observée « tombe » dans cet intervalle.

9.4.2 Les facteurs de dépendance de la longueur de l'intervalle de pari (IP)

La longueur de l'IP est $2u_\alpha\frac{\sigma}{\sqrt{n}}$

- la longueur dépend de α
  Si $\alpha' < \alpha$, la longueur de $IP_{1-\alpha'}$ est supérieure à la longueur de $IP_{1-\alpha}$

  Exemple

  $\alpha = 0,05 \Rightarrow u_{0,05} = 1,96$

  $\alpha = 0,01 \Rightarrow u_{0,01} = 2,57$

- la longueur dépend de *n*

  La longueur de $\mathrm{IP}_{1-\alpha}$ décroît avec *n*. C'est le reflet du fait connu selon lequel les fluctuations d'échantillonnage s'estompent avec *n*

  Exemple

  Dans le cas ci-dessus, si on remplace *n* = 64 par *n* = 6400, on obtient $\mathrm{IP}_{0,95} = [0{,}19 \,;\, 0{,}21]$

  Remarque

  Pour réduire dans un rapport 2 la longueur de l'IP, il faut un échantillon 4 fois plus grand (2^2).

9.4.3 L'intervalle de pari d'une variable aléatoire

Ce que l'on a dit pour une moyenne observée peut s'envisager pour une variable *X* quelconque dont on connaît la distribution.

L'IP de niveau 1 - α est l'intervalle [*a*, *b*] tel que $Pr(a \leq X \leq b) = 1 - \alpha$.

Exemple :

X ~ N(0, 1)

$\mathrm{IP}_{1-\alpha} = [-u_\alpha \,;\, u_\alpha]$

Une valeur numérique à retenir :

pour une variable aléatoire normale centrée réduite $\mathrm{IP}_{0,95} = [-1{,}96 \,;\, 1{,}96]$

Résumé du chapitre 9

1. **Propriétés de la moyenne arithmétique M_n d'une variable aléatoire X, moyenne calculée sur n unités statistiques :**

moyenne « vraie » de M_n = moyenne « vraie » de X

variance « vraie » de M_n = $\dfrac{\text{variance « vraie » de } X}{n}$

2. **Théorème central limite**
Si X a pour moyenne « vraie » μ, pour variance « vraie » σ^2, M_n est, lorsque n est suffisamment grand ($n \geq 30$, ou $n\Pi$ et $n(1 - \Pi) \geq 5$), à peu près distribuée comme une variable normale de moyenne « vraie » μ et de variance « vraie » σ^2/n, ce que l'on écrit :

$$M_n \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) \text{ ou } \frac{M_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathrm{N}(0, 1)$$

3. **Intervalle de pari (I. P.)**
Lorsque les conditions ci-dessus sont satisfaites, l'intervalle

$$\mathrm{IP}_{1-\alpha} = \left[\mu - u_\alpha \frac{\sigma}{\sqrt{n}} ; \mu + u_\alpha \frac{\sigma}{\sqrt{n}}\right]$$

a la propriété suivante :

$$Pr(M_n \in \mathrm{IP}_{1-\alpha}) = 1 - \alpha$$

4. **Cet intervalle s'appelle intervalle de pari (I. P.) de niveau 1-α, ou intervalle de pari au risque α.**

Chapitre 10 : Estimation - Intervalle de confiance

10.1 Introduction

Le problème de l'estimation statistique est le suivant : on cherche à connaître les valeurs de certaines caractéristiques d'une variable aléatoire grâce à des observations réalisées sur un échantillon. Un grand nombre de problèmes statistiques consistent en la détermination de la moyenne « vraie », sur la base d'observations réalisées sur un échantillon. Cependant, on peut aussi chercher à connaître les valeurs d'autres caractéristiques, comme par exemple les variances (exemple c. ci-dessous).

Exemples :

a. quelle est la fréquence de survenue de tel type de cancer chez les souris ?
b. quelle est la vraie valeur de la glycémie de ce patient ?
c. quelle est la variance de la glycémie mesurée chez ce patient ?

Il est bien sûr impossible de répondre à ces questions au sens strict.

On y apporte généralement deux types de réponses :

1. On produit une valeur qui nous semble être la meilleure possible : on parle alors d'**estimation ponctuelle**.
2. On produit un intervalle de valeurs possibles, compatibles avec les observations. C'est la notion d'**intervalle de confiance** ou d'**estimation par intervalle**.

Dans la suite on note X la variable aléatoire dont on cherche à estimer une caractéristique, aussi appelée paramètre, dont la valeur est notée θ. Par exemple le paramètre peut être la glycémie, et sa valeur celle du patient considéré.

10.2 Estimation ponctuelle

10.2.1 Définition

A partir d'un échantillon $(X_1, X_2, ..., X_n)$ de la variable aléatoire X, on construit une nouvelle variable aléatoire $t(X_1,X_2, ..., X_n)$ dont les réalisations « se rapprochent » de la valeur θ. Cette nouvelle variable est appelée **estimateur** de θ. Pour simplifier, cette variable $t(X_1, X_2, ..., X_n)$ est notée T_n ou T.

Par exemple $t(X_1, X_2, ..., X_n) = M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ « se rapproche » de l'espérance de X (voir chapitre 9).

C'est un estimateur naturel de $E[X]$.

10.2.2 Propriétés

Les estimateurs sont des fonctions des échantillons : ce sont donc des variables aléatoires qui possèdent une densité de probabilité, et le plus souvent, une moyenne (espérance mathématique) et une variance. Ces deux grandeurs permettent de comparer, dans une certaine mesure, les estimateurs entre eux.

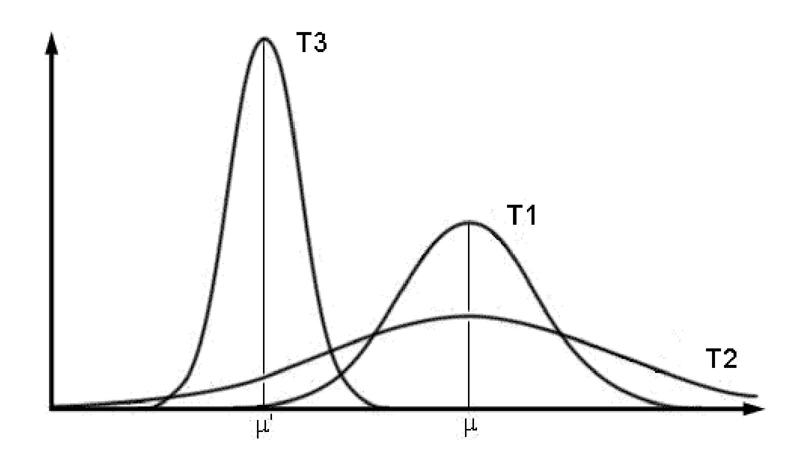


Figure 8 : densité de probabilité de 3 estimateurs T1, T2 et T3

La figure 8 représente les densités de probabilité de 3 estimateurs T1, T2 et T3 d'une moyenne μ.

10.2.2.1 Biais

On voit sur la figure 8 que T1 et T2 sont centrés autour de μ, tandis que T3 a pour moyenne μ' inférieure à μ. Cette notion est définie plus précisément de la manière suivante :

Le **biais** d'un estimateur, noté B(T), est la différence moyenne entre sa valeur et celle de la quantité qu'il estime. On a :

$B(T) = E(T - \theta) = E(T) - \theta$

Ici, on a : $B(T1) = E(T1 - \mu) = E(T1) - \mu = 0$

de même : $B(T2) = 0$

mais : $B(T3) = E(T3 - \mu) = E(T3) - \mu = \mu' - \mu < 0$

On dit que T1 et T2 sont des estimateurs sans biais de μ, et que T3 est un estimateur biaisé de μ.

10.2.2.2 Variance

La variance d'un estimateur est définie de la manière usuelle :

$var(T) = E[T - E(T)]^2$

Si deux estimateurs sont sans biais, le meilleur est celui qui a la variance la plus petite : en effet, ses valeurs sont « en moyenne » plus proches de la quantité estimée.

Par exemple, sur la figure ci-dessus, on voit que $var(T1) < var(T2)$. On peut donc conclure que T1 est un meilleur estimateur de μ que T2.

Quand des estimateurs sont biaisés, en revanche, leur comparaison n'est pas aussi simple : un estimateur peu biaisé, mais de variance très faible, pourrait même, en pratique, être préféré à un estimateur sans biais, mais de variance grande.

10.2.2.3 Erreur quadratique moyenne

L'erreur quadratique moyenne est une grandeur permettant de comparer des estimateurs entre eux, qu'ils soient biaisés ou sans biais. Elle est définie de la manière suivante :

$EQM(T) = E[(T - \theta)^2]$

On démontre facilement qu'on peut relier l'erreur quadratique moyenne, l'espérance et la variance d'un estimateur par l'expression suivante :

$EQM(T) = var(T) + [E(T) - \theta]^2 = var(T) + B(T)^2$

En particulier, l'erreur quadratique moyenne des estimateurs sans biais est égale à leur variance.

Lorsqu'on compare deux estimateurs, on considère que le meilleur est celui qui présente l'erreur quadratique moyenne la plus faible.

10.2.3 Exemple

On a souvent utilisé, dans ce cours, les quantités m, moyenne observée, et s^2, variance observée. La variable aléatoire moyenne arithmétique, notée M_n, a été étudiée au chapitre 8. De la même manière, étudions la variable aléatoire variance S_n^2, définie par :

$$S_n^2 = \frac{n}{n-1}[M_{2,n} - M_n^2]$$

où $M_{2,n}$ est la variable aléatoire « moyenne arithmétique de X^2 ».

On va calculer $E(S_n^2)$. On rappelle que si U est une variable aléatoire, la variable moyenne arithmétique définie sur U a les propriétés suivantes :

$$E(M_{U,n}) = E(U) \quad (1) \text{ et } var(M_{U,n}) = \frac{1}{n}var(U) \quad (2)$$

On a par ailleurs :

$var(U) = E(U^2) - [E(U)]^2$ et donc $E(U^2) = var(U) + [E(U)]^2$ (3).

On peut maintenant calculer $E(S_n^2)$. Soit X une variable aléatoire d'espérance $E(X) = \mu$ et de variance $var(X) = \sigma^2$. On a :

$$E(S_n^2) = \frac{n}{n-1}[E(M_{2,n}) - E(M_n^2)]$$

Mais $E(M_{2,n}) = E(X^2) = \sigma^2 + \mu^2$ d'après (1) et (3),

et $E(M_n^2) = var(M_n) + [E(M_n)]^2 = \frac{\sigma^2}{n} + \mu^2$ d'après (3), (2) et (1),

et finalement : $$E(S_n^2) = \frac{n}{n-1}\left[\sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2\right] = \sigma^2$$.

S_n^2 est donc un estimateur sans biais de σ^2.

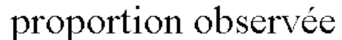
10.3 - Estimation par intervalle - Intervalle de confiance

10.3.1 Exemple d'une proportion

L'idée directrice est la suivante : on souhaite associer à une valeur observée p un intervalle appelé INTERVALLE DE CONFIANCE qui ait « de bonnes chances » de contenir la valeur « vraie » Π de la proportion. Que signifie de « bonnes chances » ? Si l'on effectue un grand nombre de fois l'expérience - chaque expérience produisant un pourcentage observé p - on construit autant d'intervalles de confiance. On voudrait qu'un grand nombre de ces intervalles contienne la valeur « vraie » Π. Par exemple que 95 % des intervalles en gros contiennent Π. On parlera alors d'intervalle de confiance DE NIVEAU 0,95 ou d'intervalle de confiance AU RISQUE 0,05. On considérera généralement des intervalles de confiance de niveau 1-α. La valeur α sera alors le risque - ou la probabilité - pour qu'un intervalle de confiance ne contienne pas la proportion « vraie » Π.

DE FACON GENERALE, L'INTERVALLE DE CONFIANCE AU RISQUE α D'UNE VALEUR QUE L'ON CHERCHE A ESTIMER EST UN INTERVALLE QUI CONTIENT AVEC UNE PROBABILITE 1 - α LA VALEUR CHERCHEE ; IL S'AGIT D'UN INTERVALLE QUE L'ON DEVRA ETRE EN MESURE DE CONSTRUIRE A L'ISSUE D'UNE EXPERIENCE PORTANT SUR UN ECHANTILLON.

Comment construire de tels intervalles ? C'est facile graphiquement.



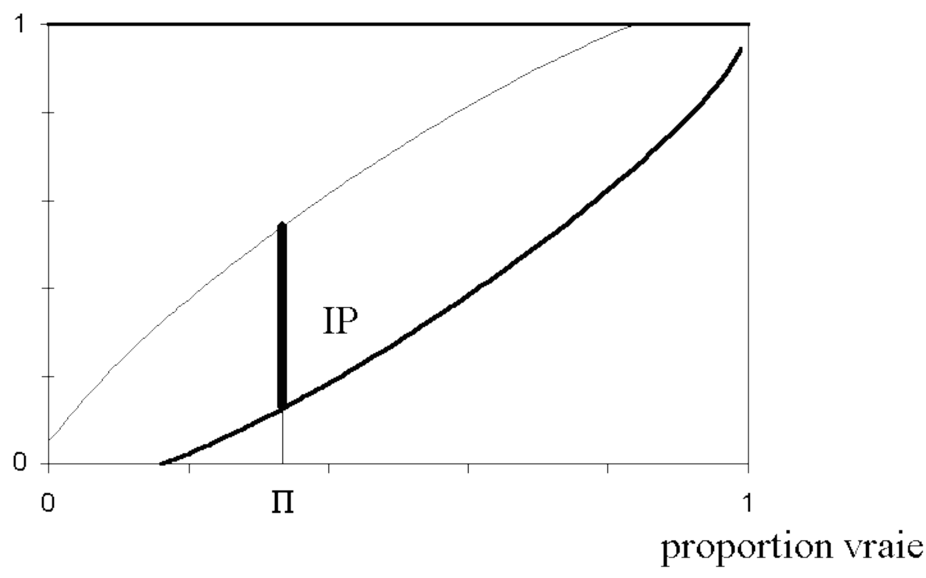


Figure 9

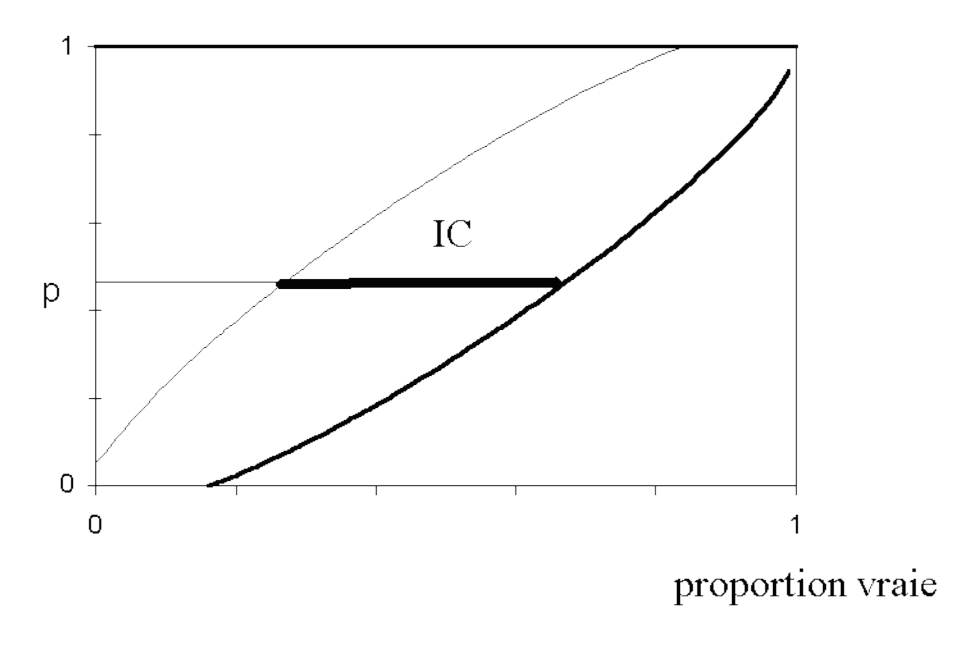


Figure 10

Considérons la figure 9. On a porté en abscisses une échelle 0-1 de mesure de proportions « vraies », en ordonnées une échelle de mesure de proportions observées. Donnons nous une valeur de proportion « vraie » ; on sait associer à cette valeur un intervalle de pari de niveau 0,95 de la proportion observée que l'on est susceptible d'obtenir au cours d'une expérimentation conduite sur *n* individus. Cet intervalle de pari peut être représenté sur

l'échelle verticale. Si l'on opère cette représentation pour toutes les valeurs possibles d'une proportion « vraie », on obtient un domaine limité par les deux courbes représentées sur la figure.

Considérons alors un problème mettant en jeu une proportion « vraie », Π. Supposons que nous fassions un ensemble d'expériences, chaque expérience portant sur n individus étant productive d'une valeur de proportion observée p. On peut associer à chacune de ces expériences un point de coordonnées (Π, p) sur la figure 9. Compte tenu de la construction précédente, on peut affirmer que ces points appartiendront 95 fois sur cent (c'est-à-dire dans 95 % des expériences) au domaine limité par les deux courbes, et ceci quelle que soit la valeur de Π.

Maintenant supposons qu'une expérience unique ait été réalisée, produisant une valeur de proportion, p. Le problème est, sur la base de cette valeur, de définir un intervalle ayant de bonnes chances de contenir la valeur inconnue de la proportion « vraie ». La solution, immédiate, est fournie par la figure 10. Il suffit de trancher le domaine limité par les deux courbes DANS L'AUTRE SENS. Cet intervalle contiendra 95 fois sur cent la véritable valeur de la proportion.

Ainsi, si on adopte cette stratégie de construction, on aura pour chaque valeur observée p un intervalle qui contiendra Π avec la probabilité 0,95.

Le problème est résolu. Maintenant, ce qui est simple sur un dessin est compliqué en termes de calcul et il existe des tables d'intervalles de confiance et des formules toutes faites permettant de former des intervalles de confiance approchés.

10.3.2 Intervalle de confiance approché d'une proportion « vraie »

On montre qu'une bonne approximation de l'intervalle de confiance de niveau 1 - α de Π, fondé sur la valeur observéep, p étant calculée sur n individus, est donnée par l'intervalle ci-dessous :

$$\mathrm{IC}_{1-\alpha} = \left[p - u_\alpha \sqrt{\frac{p(1-p)}{n}}\ ;\ p + u_\alpha \sqrt{\frac{p(1-p)}{n}}\right]$$

Notons Π_{min} et Π_{max} les bornes de cet intervalle.

Cette approximation n'est jugée satisfaisante que sous les CONDITIONS DE VALIDITE suivantes : $n\Pi_{min} \geq 5$, $n(1-\Pi_{max}) \geq 5$

LORSQUE LES CONDITIONS DE VALIDITE NE SONT PAS REMPLIES, IL FAUT AVOIR RECOURS A DES TABLES (hors programme).

Exemple : n = 100, α = 0,05, p = 0,12

$$\mathrm{IC}_{0,95} = \left[0,12 - 1,96\sqrt{\frac{0,12 \times 0,88}{100}}\ ;\ 0,12 + 1,96\sqrt{\frac{0,12 \times 0,88}{100}}\right] = [0,06\ ;\ 0,18]$$

conditions de validité

$100 \times 0{,}06 = 6 \geq 5$.

$100 \times (1 - 0{,}18) = 82 \geq 5$.

10.3.3 Intervalle de confiance approché d'une moyenne « vraie » (variable continue)

De même, il existe une expression approchée pour l'intervalle de confiance de niveau 1 - α d'une moyenne « vraie » μ, intervalle fondé sur la valeur observée *m* obtenue après une expérience portant sur *n* individus. Le calcul de cet intervalle suppose en outre le calcul de la variance observée s^2. L'expression est la suivante :

$$IC_{1-\alpha} = \left[m - u_\alpha \frac{s}{\sqrt{n}} \; ; \; m + u_\alpha \frac{s}{\sqrt{n}} \right]$$

L'approximation ci-dessus n'est jugée satisfaisante que sous la

CONDITION DE VALIDITE : $n \geq 30$.

Lorsque cette condition n'est pas remplie, on ne sait plus former d'intervalle de confiance sauf si l'on peut supposer que la variable primitive *X* d'intérêt est normale.

Si la variable étudiée est NORMALE, alors, et sans autre condition de validité, un intervalle de confiance de niveau 1 - α a pour expression :

$$IC_{1-\alpha} = \left[m - t_\alpha \frac{s}{\sqrt{n}} \; ; \; m + t_\alpha \frac{s}{\sqrt{n}} \right]$$

où t_α est associé à une nouvelle distribution, dite de Student, à (*n*-1) degrés de liberté (voir section 7.2.3). La notationt_α s'apparente à la notation u_α et est explicitée table A.6.

Remarque (pour une variable normale encore)

Si la variance « vraie » de la variable étudiée, σ^2, est connue, l'intervalle de confiance a la forme suivante :

$$IC_{1-\alpha} = \left[m - u_\alpha \frac{\sigma}{\sqrt{n}} \; ; \; m + u_\alpha \frac{\sigma}{\sqrt{n}} \right]$$

10.3 - Estimation par intervalle - Intervalle de confiance

10.3.4 - Applications

L'intervalle de confiance exprime fondamentalement, comme son nom l'indique, la confiance que l'on peut attribuer à un résultat expérimental.

IDEALEMENT TOUT PROBLEME D'ESTIMATION DEVRAIT ETRE PRODUCTIF D'UN INTERVALLE DE CONFIANCE. Ne donner qu'une estimation ponctuelle masque l'incertitude qui accompagne tout résultat.

Exemple : supposons qu'étudiant la fréquence d'un événement, on ait obtenu une fréquence observée *p* égale à 0,12.

Supposons que cette valeur ait été obtenue sur la base de 8 individus (l'événement étudié s'est donc réalisé une fois). On peut lire dans une table spécialisée que l'intervalle de confiance de la fréquence « vraie » est, au risque 0,05 [0,003 ; 0,527]. Cela signifie que cette valeur observée de 12 % sur si peu d'individus ne fait qu'indiquer ceci : la fréquence « vraie » se situe dans le domaine 3 ‰, 52,7 %.

Supposons que cette même valeur 12 % ait été obtenue sur la base de 100 individus (l'événement étudié s'est réalisé 12 fois au cours des 100 essais). L'intervalle de confiance associé est alors proche de [0,06 ; 0,18]. Sur la base de cette valeur 12 %, on est maintenant en mesure d'affirmer, acceptant toujours un risque d'erreur de 5 pour cent, que la fréquence « vraie » se situe dans le domaine 6 %, 18 %, domaine beaucoup plus étroit que le précédent.

De façon générale, la longueur de l'intervalle de confiance indique la précision obtenue. Les deux exemples qui suivent montrent l'usage que l'on peut en faire.

10.3.4.1 Précision d'un sondage

Supposons que l'on s'apprête à réaliser un sondage pour estimer la prévalence d'une maladie, c'est-à-dire la proportion de la population atteinte par cette maladie à la date du sondage. On souhaite un résultat précis, c'est-à-dire que l'on souhaite par exemple que l'intervalle de confiance résultant ait une longueur au plus égale à 0,04, avec un risque d'erreur de 5 %.

On remarque que la longueur de l'intervalle de confiance ne dépend que d'une seule grandeur contrôlable, le nombre d'individus. La question est donc : combien d'individus faut-il inclure dans le sondage ?

Ce problème est simple, puisque la longueur de l'intervalle de confiance s'établit à :

$2 \times 1{,}96 \sqrt{\frac{p(1-p)}{n}}$ qu'on arrondit ici à $4\sqrt{\frac{p(1-p)}{n}}$

L'effectif de l'échantillon devra donc être au moins 10000 $p(1 - p)$.

Toutefois, cet effectif dépend de *p*, inconnu avant l'expérience. L'usage de ces calculs supposera donc que l'on ait une idée du résultat attendu, grâce à un sondage exploratoire par exemple ou grâce à une connaissance préalable du phénomène étudié.

De façon générale, si l'on souhaite obtenir un intervalle de confiance d'une proportion de longueur 2*i*, il est nécessaire d'inclure un nombre d'individus au moins égal à :

$4\frac{p(1-p)}{i^2}$ au risque 0,05 (ou $u_\alpha^2\frac{p(1-p)}{i^2}$ au risque α)

REMARQUE : Lorsque le sondage est réalisé, un intervalle de confiance lui est associé. Dans le langage courant, les instituts de sondage nomment ces intervalles de confiance des FOURCHETTES.

10.3.4.2 Précision d'une moyenne

Dans le cas où l'on s'intéresse à la moyenne « vraie » d'une variable quantitative, on peut effectuer le même type de calcul. Pour obtenir un intervalle de confiance de longueur 2*i*, il faut inclure un nombre d'individus au moins égal à :

$$n = u_{\alpha}^{2}\frac{s^{2}}{i^{2}}$$

L'exploitation de ce calcul nécessite ici une connaissance, même approximative, de la variance de la variable étudiée pour se donner a priori s^2- ou mieux σ^2.

Exemple très important : les problèmes de dosage.

Soit à doser la glycémie ; on a devant soi un échantillon de sang. Quelle est la concentration en glucose ? Si on fait plusieurs dosages, on va obtenir plusieurs résultats. Cela est dû, non à la variabilité de la glycémie, mais aux erreurs analytiques. On assimile la glycémie « vraie » à la moyenne « vraie » de la variable aléatoire « résultat du dosage ». Supposons que l'on connaisse la variance des résultats, car on connaît bien la technique analytique. Par exemple, σ = 10 mg.l^{-1}. Supposons en outre que les résultats expérimentaux soient distribués normalement.

Si on effectue un dosage donnant 90 mg.l^{-1}, on a pour intervalle de confiance approché (σ étant connu) :

$IC_{0,95}$ = [90 - 2σ ; 90 + 2σ] = [70 ; 110] soit un intervalle de longueur 40.

Si on effectue deux dosages donnant 90 et 96 mg.l^{-1}, on a

$$IC_{0,95} = \left[93 - 2\frac{\sigma}{\sqrt{2}} \,;\, 93 + 2\frac{\sigma}{\sqrt{2}}\right] = [78,9 \,;\, 107,1]$$

soit un intervalle d'amplitude 28,2.

Si l'on effectue trois dosages donnant 90, 96 et 93 mg.l^{-1} on a

$$IC_{0,95} = \left[93 - 2\frac{\sigma}{\sqrt{3}} \,;\, 93 + 2\frac{\sigma}{\sqrt{3}}\right] = [81,5 \,;\, 104,5]$$

soit un intervalle d'amplitude 23,0.

Ces calculs objectivent le fait bien connu selon lequel la répétition des dosages permet d'atténuer les conséquences des erreurs expérimentales. Certains dosages - certaines mesures (tension artérielle) - sont répétés avant qu'une valeur soit indiquée.

Chapitre 11 - Les tests d'hypothèses. Principes

Les tests d'hypothèses sont fondés sur les intervalles de pari.

11.1 Un exemple concret (emprunté à Schwartz)

Une variété de souris présente des cancers spontanés avec un taux (une fréquence ou proportion dans la population) constant bien connu, $\varphi = 20$ %. On se demande si un traitement donné modifie ce taux (en plus ou en moins), c'est-à-dire est actif. Pour répondre à cette question on procède à une expérience sur 100 souris ; il s'agira, au vu du pourcentage observé p d'animaux cancéreux, de dire si le traitement est actif. Il n'est pas possible de répondre au sens strict à cette question.

Supposons que le traitement soit sans effet ; alors chaque souris traitée aura toujours 20 chances sur 100 de devenir cancéreuse. Mais le pourcentage de souris cancéreuses, calculé sur un échantillon de 100 souris sera soumis aux fluctuations d'échantillonnage que l'on a étudiées. Le pourcentage observé (moyenne observée) pourra prendre a priori, c'est-à-dire avant expérience, plusieurs valeurs, même si les valeurs voisines de 0,2 sont les plus probables. Des valeurs de 0 ou 100 % pourraient même être observées. Ainsi même si le pourcentage observé est très différent de 20 %, il est possible que le traitement soit sans effet.

Supposons maintenant que le traitement soit actif ; la probabilité de cancer pour chaque souris (ou la proportion « vraie » de souris cancéreuses dans une population fictive de souris traitées) est φ_1, différente de 0,2. Encore à cause des fluctuations d'échantillonnage, on pourra très bien, peut être de façon peu probable, obtenir une fréquence observée égale à 20 %. Ainsi même si le pourcentage observé est 20 %, il est possible que le traitement soit actif.

On ne peut donc répondre avec certitude à la question posée.

Pourtant ne pas répondre serait renoncer à considérer tous les problèmes liés à la variabilité, c'est-à-dire à « tous » les problèmes biologiques. Alors on répondra, mais en acceptant un risque d'erreur. Répondre correspond à la démarche que chacun adopterait ; par exemple, déclarer le traitement actif si le taux observé de cancers après traitement s'écarte « nettement » de 20 %. C'est le sens que l'on peut donner à ce « nettement » qui est le fondement du principe des tests.

Dans le cas étudié, on aurait tendance à s'y prendre de la façon suivante. Deux hypothèses sont en présence :

- le traitement est inactif,
- le traitement est actif.

La première hypothèse est plus « fine » que la seconde car elle porte en elle une interprétation numérique : le pourcentage « vrai » de souris cancéreuses parmi les souris traitées est 0,2 - l'autre hypothèse indiquant seulement que ce pourcentage est différent de

0,2 ; ce qui est plus vague. Supposons alors vraie l'hypothèse la plus fine. Il devient possible de faire des déductions : sachant ce qui se passe au niveau de la population des souris traitées on peut en déduire ce qui se passera au niveau d'un échantillon. En particulier, on sait construire les intervalles de pari centrés de niveau 1 - α pour la fréquence observée.

Par exemple, prenant $\alpha = 0{,}05$ et $n = 100$ souris, on obtient $IP_{0,95} = [0{,}12 ; 0{,}28]$

Cela signifie, rappelons-le, que si $\varphi = 0{,}2$ (fréquence supposé « vraie »), 95 % des valeurs des moyennes observées calculées sur 100 individus appartiendront à l'intervalle
[0,12 ; 0,28].

On adopte alors la stratégie suivante : si la valeur observée de la fréquence de souris cancéreuses parmi les 100 traitées appartient à cet intervalle, on considère que cette valeur est compatible avec les fluctuations d'échantillonnage et l'activité du traitement n'est pas prouvée. Si la valeur observée n'appartient pas à cet intervalle, le traitement sera considéré comme actif. Dans ce dernier cas le raisonnement est le suivant. Cet événement (la fréquence observée est à l'extérieur de l'intervalle de pari) avait moins de 5 chances sur 100 de se produire et pourtant il s'est produit ; donc je ne crois plus à l'hypothèse qui m'a permis de déduire ces 5 % de chances.

Remarque : reformulation des calculs

Notons *p* la proportion observée de souris traitées développant un cancer, sur les *n* souris traitées.

Le résultat du test sera de conclure ou non à l'activité du traitement selon que $p \notin \text{ ou } \in IP_{1-\alpha}$ c'est-à-dire :

$$p \notin \text{ ou } \in \left[\varphi_0 - u_\alpha\sqrt{\frac{\varphi_0(1-\varphi_0)}{n}} \; ; \; \varphi_0 + u_\alpha\sqrt{\frac{\varphi_0(1-\varphi_0)}{n}}\right]$$

où φ_0 est la proportion hypothétique (0,2 dans l'exemple) et u_α la borne de l'intervalle de pari au risque α de *p*.

On suppose ici que les conditions du théorème central limite sont satisfaites. On conclut donc selon que :

$$p - \varphi_0 \notin \text{ ou } \in \left[-u_\alpha\sqrt{\frac{\varphi_0(1-\varphi_0)}{n}} \; ; \; u_\alpha\sqrt{\frac{\varphi_0(1-\varphi_0)}{n}}\right]$$

ou encore selon que

$$\frac{p - \varphi_0}{\sqrt{\frac{\varphi_0(1-\varphi_0)}{n}}} \notin \text{ ou } \in [-u_\alpha \; ; \; u_\alpha]$$

On reconnaît dans la dernière expression l'intervalle de pari $\mathrm{IP}_{1-\alpha}$ d'une variable aléatoire N(0, 1), intervalle indépendant de l'expérience projetée.

C'est comme cela que l'on abordera généralement les tests ; on cherchera à construire une variable aléatoire dont on connaisse, si l'hypothèse fine est vraie, la distribution, pour pouvoir construire un intervalle de pari ; ici il s'agirait de la variable aléatoire Z déduite de la variable aléatoire moyenne arithmétique selon :

$$Z = \frac{P_n - \varphi_0}{\sqrt{\frac{\varphi_0(1-\varphi_0)}{n}}}$$

avec $\varphi_0 = 0{,}2$ (transcription de l'hypothèse).

Une telle variable aléatoire s'appelle usuellement « paramètre » du test et est notée conventionnellement Z. Ici on sait que $Z \sim N(0, 1)$ et l'on construit l'intervalle de pari de niveau $1 - \alpha$ pour Z. Par exemple avec $\alpha = 0{,}05$, $\mathrm{IP}_{0,95} = [-1{,}96 ; 1{,}96]$.

Puis on réalise l'expérience ce qui permet d'obtenir p, valeur observée de P_n, donc une valeur observée de Z, notée u :

$$z = \frac{p - \varphi_0}{\sqrt{\frac{\varphi_0(1-\varphi_0)}{n}}}$$

On pourrait alors s'exprimer comme ceci (une terminologie plus précise sera indiquée plus loin) :

- si $z \in \mathrm{IP}_{0,95}$ on ne peut pas dire que le traitement est actif
- si $z \notin \mathrm{IP}_{0,95}$ le traitement est actif.

Nous allons, à la lumière de cet exemple, énumérer les étapes de mise en œuvre d'un test et revenir sur différents aspects (sens de α par exemple) avant de donner d'autres exemples de tests usuels

11.2 - Principe général des tests d'hypothèses

La mise en œuvre d'un test statistique nécessite plusieurs étapes.

11.2.1 Les étapes de mises en œuvre

Etape 1

Avant le recueil des données.

Définir avec précision les deux hypothèses en présence H_0 et H_1. H_0 et H_1 jouent toujours des rôles dissymétriques.
Le plus souvent, une des hypothèses est précise, ou fine. Elle engage une égalité généralement ; c'est elle qui sera H_0 et on l'appellera **hypothèse nulle**,

H_0 : hypothèse nulle

Exemple : la fréquence « vraie » d'apparition du cancer chez les souris traitées est 0,2, ce qui se transcrit par $\phi = 0,2$ (plus généralement $\phi = \phi_0$).

Le principe des tests est d'admettre cette hypothèse H_0 sauf contradiction flagrante entre ses conséquences et les résultats expérimentaux.

L'autre hypothèse est toujours plus vague ; **elle regroupe toutes les hypothèses, hormis H_0**. C'est H_1 et on l'appellera **hypothèse alternative**,

H_1 : hypothèse alternative

Exemple : la fréquence « vraie » d'apparition du cancer chez les souris traitées est différente de 0,2, qui se transcrit par $\phi \neq 0,2$ (généralement $\phi \neq \phi_0$).

Remarque : la formulation de ces hypothèses nécessite généralement une traduction et une simplification du problème médical sous-jacent.

Etape 2

Avant le recueil des données.

On suppose que H_0 est vraie et on cherche à définir une variable aléatoire (ou paramètre) dont on connait alors la distribution. En d'autres termes, on cherche à construire une fonction des données à venir dont on connait la distribution si H_0 est

vraie. Soit Z cette variable aléatoire.

Exemple :

$$Z = \frac{P_n - \varphi_0}{\sqrt{\frac{\varphi_0(1-\varphi_0)}{n}}} \sim N(0, 1)$$

Si possible, vérifier les conditions de validité.

Etape 3

Avant le recueil des données.

Choisir un seuil. Typiquement α = 0,05 (une quasi obligation en pratique)

Construire un intervalle de pari (pour le paramètre Z) de niveau 1 - α, noté $\mathrm{IP}_{1-\alpha}$.
Rappelons qu'il s'agit d'un intervalle tel que si H_0 est vraie, alors

$$P(Z \in \mathrm{IP}_{1-\alpha}) = 1 - \alpha$$

Exemple : $\mathrm{IP}_{1-\alpha}$ pour Z ci-dessus = [-1,96 ; 1,96]

Définition : l'extérieur de l'intervalle de pari $\mathrm{IP}_{1-\alpha}$ s'appelle **région critique du test au seuil α**.

Etape 4

Avant le recueil des données.

Définir la règle de décision. Les données vont permettre de calculer une valeur de Z, que l'on note z.

Exemple :

$$z = \frac{p_{\text{réellement observé}} - \varphi_0}{\sqrt{\frac{\varphi_0(1-\varphi_0)}{n}}}$$

Alors décider que :

- si z appartient à la région critique, remettre en cause H_0,
  la **rejeter**, et conclure **H_1 est vraie**, ou dire : « au risque α, H_0 est rejetée ».

- si z n'appartient pas à la région critique, mais à l'intervalle de pari $IP_{1-\alpha}$, dire que l'on ne conclut pas, ou dire que l'on ne rejette pas l'hypothèse nulle H_0.

Etape 5

Recueil des données

Réaliser l'expérience. On recueille les données $x_1, ..., x_n$; calculer z et conclure.
Si non fait à l'étape 2, vérifier les conditions de validité.

Etape 6

Interprétation des résultats

Cette étape concerne l'interprétation des résultats en des termes compatibles avec le problème médical initialement soulevé, et concerne en particulier le problème de la causalité. Ce point sera détaillé au chapitre 15.

Exemple : dans le cas des souris, et en cas de conclusion au rejet de l'hypothèse nulle, la question serait de savoir si ce rejet exprime véritablement une activité du traitement.

11.2.2 - Justification de la règle de décision. Choix de α

11.2.2.1 Interprétation de α

On a déjà vu une interprétation de α avec l'exemple des souris. De façon générale, α est la probabilité pour que la valeur observée - ou calculée - z appartienne à la région critique si H_0 est vraie. Si cet événement se réalise, on rejette H_0. Cela ne se justifie que si α est petit car alors on dit : voilà un événement qui avait 100×α % chances de se réaliser (5 % par exemple) - donc peu de chances - et qui pourtant s'est réalisé : les résultats ne sont pas conformes à l'hypothèse ⇒ α doit être petit.

Une autre interprétation de α montre encore mieux que α doit être petit. A nouveau, lorsque H_0 est vraie, la probabilité d'obtenir un résultat z dans la région critique est α. Mais alors on dit « H_1 est vraie ». Donc

⇒ α = « probabilité » de conclure H_1 alors que H_0 est vraie

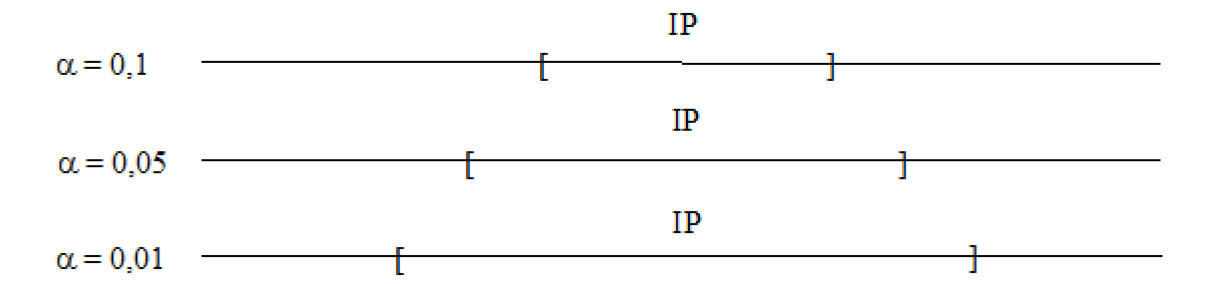
C'est un risque d'erreur qu'il convient de situer dans des valeurs acceptables (petites).

Cette valeur α s'appelle **RISQUE DE PREMIERE ESPECE**.

Cela veut dire que sur un grand nombre d'expériences, en admettant α, on conclura à tort dans 100×α % des cas (5 % des cas par exemple). Pourquoi alors ne pas choisir un α microscopique ?

11.2.2.2 Effet d'un changement de valeur de α

Les intervalles de pari croissent lorsque leur niveau augmente, c'est-à-dire lorsque α diminue.



Donc, toutes choses égales par ailleurs, la région critique diminue lorsque α décroît. Donc on rejette moins fréquemment H_0.

A vouloir commettre moins d'erreurs, on conclut plus rarement.

On s'expose donc à un autre risque : celui de ne pas conclure alors qu'il le faudrait car H_0 est fausse. A la limite, si on se fixe $\alpha = 0$, on ne conclut jamais, H_0 n'est jamais rejetée.

Prendre une décision, c'est accepter un risque.

Pour finir avec ce problème de α il faut retenir :

- La valeur de α doit être fixée a priori : jamais en fonction des données
- Pire que cela, on choisit la valeur $\alpha = 0,05$ qui est un compromis entre le risque de conclure à tort et la faculté de conclure, compromis adopté par l'ensemble de la communauté scientifique.

11.2.3 - Justification des conclusions du test. Puissance d'un test

On comprend maintenant la partie de la règle de décision conduisant au rejet de H_0 lorsque la valeur calculée du paramètre n'appartient pas à l'intervalle de pari. On a par ailleurs indiqué (voir l'étape 4 de mise en œuvre des tests) que lorsque la valeur calculée du paramètre appartient à l'intervalle de pari, c'est-à-dire lorsque les résultats expérimentaux ne sont pas contradictoires avec l'hypothèse nulle, on s'exprime avec beaucoup de précautions oratoires puisqu'on demande de dire : « on ne conclut pas » ou « on ne rejette pas l'hypothèse nulle ». Pourquoi ne pas affirmer plus directement « l'hypothèse nulle est vraie » ?

Premier élément

En faisant cela, on adopte une démarche qui s'apparente à la démarche scientifique qui consiste à admettre une théorie jusqu'à la preuve de son échec. Lorsque l'on dit « admettre » on ne signifie pas que la théorie est vraie mais qu'elle rend compte pour l'instant - jusqu'à plus ample informé - des expériences.

Exemples

- la mécanique générale admise jusqu'à la théorie de la relativité
- la mécanique céleste

Second élément

Supposons que l'on mette en parallèle les deux tests suivants :

$H_0 : \varphi = 0{,}2$	$H_0 : \varphi = 0{,}200001$
$H_1 : \varphi \neq 0{,}2$	$H_1 : \varphi \neq 0{,}200001$

Les paramètres calculés, soit

$$z = \frac{p_{\text{observée}} - \varphi_0}{\sqrt{\frac{\varphi_0(1-\varphi_0)}{n}}} \quad \text{avec } \varphi_0 = 0{,}2 \text{ ou } 0{,}200001$$

seront extrêmement voisins, donc les conclusions pratiquement toujours les mêmes.

Considérons alors une expérience au cours de laquelle $z \in IP_{0,95}$ pour les deux valeurs calculées. Peut-on conclure à la fois φ = 0,2 et φ = 0,200001 ? Pourtant on peut remarquer qu'il n'y a pas de vice de fond au niveau de la formulation des hypothèses car il existe bien une valeur « vraie », c'est-à-dire qu'il y a vraiment une hypothèse vraie du type φ = quelque chose.

On retient : **les tests ne sont pas faits pour « démontrer » H_0, mais pour la rejeter.** Cela ne veut pas dire que l'on est toujours content de rejeter H_0.

Exemples

- cas des souris traitées. Là on aimerait probablement rejeter H_0, c'est-à-dire conclure à l'activité du traitement.
- cas d'un test d'homogénéité. On vous livre un nouveau lot de souris ou des souris d'un autre élevage. Vous voulez continuer vos recherches. La première chose à faire est de tester l'hypothèse selon laquelle ces nouvelles souris sont similaires aux précédentes vis-à-vis du taux de cancer, ⇒ $H_0 : \varphi = 0{,}2$. Mais là

vous espérez bien ne pas rejeter H_0. C'est à cette condition que vous pouvez continuer.

PUISSANCE D'UN TEST

Revenons à la conclusion « l'activité du traitement n'est pas démontrée ». Sous entendu compte tenu de l'expérience effectuée. Cela n'a de sens de s'exprimer comme cela que s'il est pensable qu'une autre expérience, plus complète par exemple, puisse montrer cette efficacité si elle existe.
C'est le cas, en effet. L'aptitude d'un test à rejeter l'hypothèse nulle alors qu'elle est fausse est limitée. Précisément :

On appelle **PUISSANCE D'UN TEST** *P* la probabilité de rejeter l'hypothèse nulle, face à une hypothèse alternative, alors qu'elle est fausse.

La valeur complémentaire à 1 de cette puissance, c'est-à-dire la probabilité de ne pas rejeter l'hypothèse nulle alors que l'hypothèse alternative est vraie, s'appelle le **RISQUE DE DEUXIEME ESPECE** et se note conventionnellement β : $\beta = 1 - P$.

Le calcul de la puissance d'un test est une opération complexe. La difficulté tient essentiellement au fait que l'hypothèse alternative est vague. Pour contourner cette difficulté et apprécier plus étroitement cette notion de puissance, considérons le cas d'une hypothèse alternative fine. Par exemple, reprenant l'exemple des souris, supposons que l'hypothèse H_1 soit $\varphi = 0{,}3$, l'hypothèse H_0 restant inchangée, c'est-à-dire $\varphi = 0{,}2$. Dans ces conditions, il est possible de calculer la distribution de la proportion observée, non plus seulement sous l'hypothèse nulle, mais également sous l'hypothèse alternative. On obtient :

- sous l'hypothèse nulle ($\varphi = 0{,}2$) : $P_n \sim N\left(0{,}2, \frac{0{,}2(1-0{,}2)}{n}\right)$
- sous l'hypothèse alternative ($\varphi = 0{,}3$) : $P_n \sim N\left(0{,}3, \frac{0{,}3(1-0{,}3)}{n}\right)$

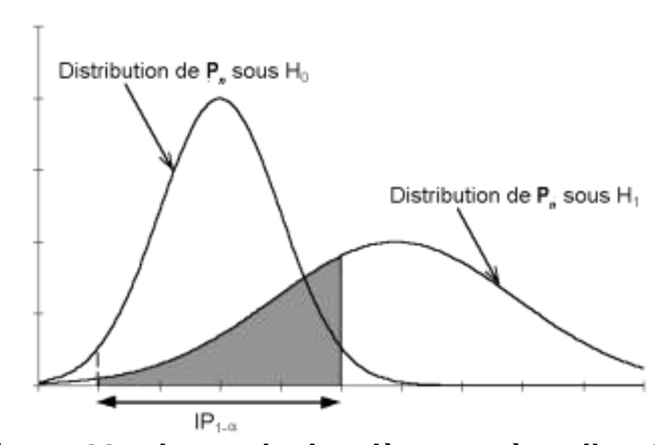


Figure 11 : risque de deuxième espèce d'un test

La figure 11 présente les deux distributions correspondantes, pour une certaine valeur de *n*. Supposons alors juste l'hypothèse H_1 ; la valeur observée *p* sera issue de la

distribution de droite, et l'on conclura à tort au non rejet de H_0 avec une probabilité égale à l'aire grisée, puisque cette aire est la probabilité pour que la valeur observée appartienne à l'intervalle de pari associé au test, sachant que cette valeur observée est gouvernée par la distribution associée à H_1. Ainsi la valeur de cette aire grisée exprime le risque de deuxième espèce β, son complémentaire à 1 la puissance du test. Supposons pour fixer les idées que la valeur de cette aire soit 0,4. Cela signifie que si les hypothèses sont φ = 0,2 et φ = 0,3, on aura « 6 chances sur dix » seulement de rejeter l'hypothèse φ = 0,2 lorsque φ sera égal à 0,3. Autrement dit, 4 fois sur dix, on sera incapable de détecter que φ vaut 0,3 et non 0,2.

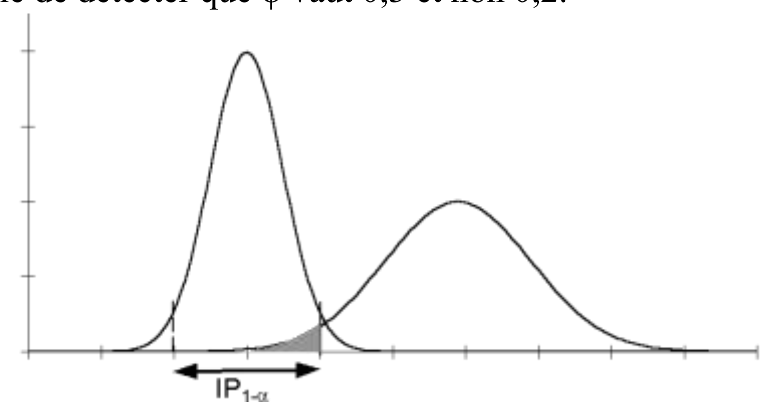


Figure 12 : risque de deuxième espèce d'un test

Par ailleurs, on perçoit que plus les hypothèses H_0 et H_1 sont contrastées (par exemple les hypothèses φ = 0,2, φ = 0,4 sont plus contrastées que les hypothèses φ = 0,2, φ = 0,3), plus les distributions de P_n sous ces deux hypothèses sont « éloignées », et plus la puissance est grande. C'est la raison pour laquelle on dit souvent que la notion de puissance est proche de la notion de pouvoir discriminant entre hypothèses.

La figure 12 reproduit les conditions de la figure 11, mais avec une valeur de n accrue. Autrement dit le même test est mis en œuvre, mais sur un nombre d'unités statistiques supérieur. On constate sur cette figure que le risque de deuxième espèce est très faible. Ce résultat est général :

TOUTES CHOSES EGALES PAR AILLEURS, LA PUISSANCE D'UN TEST AUGMENTE AVEC LA TAILLE DE L'ECHANTILLON

Remarque

Les calculs de puissance ébauchés ci-dessus, joints au résultat précédent, permettent de répondre à des questions du type :

- combien de sujets est-il nécessaire d'inclure dans un essai pour avoir de bonnes chances (9 chances sur dix par exemple) de mettre en évidence une différence entre proportions « vraies » d'au moins 0,1 ?
- si je dispose de 100 sujets, quelle différence minimum entre proportions « vraies » suis-je capable de détecter avec une probabilité de 0,9 ?

Des formules de la relation entre puissance et taille des échantillons seront données dans le chapitre 12.

Les développements ci-dessus montrent que lorsque vous n'avez pas rejeté l'hypothèse nulle, vous pouvez toujours dire que c'est un **manque de puissance du test** puisque H_0 est sans doute fausse (pensons à $\varphi = 0{,}2$ exactement). On peut donc dire qu'avec un plus grand nombre d'individus vous auriez rejeté H_0. Cela justifie l'expression « l'activité du traitement n'est pas démontrée ».
Cependant il faut être réaliste : reprenons l'exemple des souris traitées ou non traitées. Vous avez réalisé votre expérience sur un échantillon de 1000 souris. Résultat du test : non rejet de H_0 c'est-à-dire l'activité n'est toujours pas démontrée. Il n'est pas raisonnable dans ces conditions d'évoquer un manque de puissance du test ; ce résultat suggère plutôt une très faible activité du traitement, si elle existe.

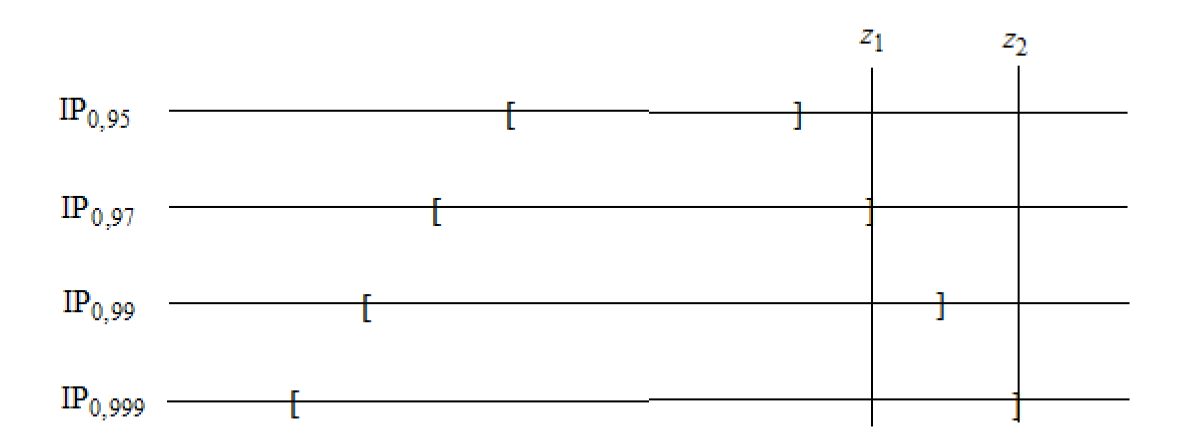
11.2.4 - Amélioration de l'interprétation du rejet de H_0

11.2.4.1 Notion de degré de signification

Supposons que l'on réalise un test au risque ou seuil $\alpha = 0{,}05$.

Considérons deux expériences conduisant au rejet de H_0, pour lesquelles on a obtenu des valeurs calculées du paramètre z_1 et z_2 représentées ci-dessous.

On aurait envie de rejeter plus fortement H_0 dans le second cas que dans le premier. En effet, considérons des intervalles de pari pour z, de niveau croissant à partir de 0,95.



On observe que z_1 est à l'extérieur des intervalles de pari jusqu'au niveau 0,97, que z_2 est à l'extérieur des intervalles de pari jusqu'au niveau 0,999. Cela signifie que, en ce qui concerne la première expérience, H_0 aurait été rejetée même si on avait limité le risque d'erreur à 1 - 0,97 = 0,03 (soit 3 %), et que, en ce qui concerne la seconde, H_0 aurait été rejetée même si on avait limité le risque d'erreur à 1 - 0,999 = 0,001 (soit 1‰). C'est ce pseudo risque d'erreur que l'on appelle **degré de signification** et qui mesure la force avec laquelle on rejette H_0.

Ce degré de signification est noté p : plus il est petit, plus confortable est le rejet.

Si l'on veut une définition plus précise :

Définition

Lorsque H_0 est rejetée, on appelle degré de signification d'un test le risque associé au plus grand intervalle de pari qui ne contient pas le paramètre calculé z.

Calcul pratique du degré de signification

On cherche dans la table la valeur de p pour laquelle $u_p = z$, u_p étant du type u_α

Exemple: $z = 2,43$.

On trouve dans la table $u_{0,02} = 2,32$ et $u_{0,01} = 2,57$

alors $p \in [0,01 ; 0,02]$

La valeur exacte ne se trouve pas dans la table : on dira $p < 0,02$. Le plus grand intervalle de pari ne contenant pas zest de niveau > 0,98, ou au risque < 0,02.

La plupart des résultats de tests s'expriment avec ce degré de signification :

- On réalise le test (avec un risque $\alpha = 0,05$)
- Si H_0 est rejetée, on calcule ou on évalue le degré de signification p
- Si H_0 n'est pas rejetée, on ne calcule pas p.

11.2.4.2 Orientation du rejet

Le rejet de H_0 correspond généralement à l'une des deux situations :

- rejet car z est trop petit (inférieur à la borne inférieure de l'intervalle de pari)
- rejet car z est trop grand (supérieur à la borne supérieure de l'intervalle de pari)

Dans le cadre de l'exemple précédent, chacune de ces situations correspond généralement à des commentaires radicalement différents. Par exemple :

z est trop petit ⇔ le traitement est efficace

z trop grand ⇔ le traitement est nuisible

11.3 - Rappels et précisions

1. LES TESTS PRENNENT EN COMPTE DES HYPOTHESES SYNTHETIQUES

   On a vu que les tests reposent sur l'énoncé de deux hypothèses exclusives. Il y a parfois beaucoup de chemin à parcourir entre la formulation d'un problème médical et sa formulation en termes statistiques. Reprenons l'exemple des souris de ce chapitre. Le problème fondamental est celui de l'activité du traitement. Cette activité peut avoir bien d'autres manifestations que la modification de la fréquence d'apparition des cancers. On peut penser à un effet portant sur l'âge de survenue de la maladie, portant sur la vitesse de développement des tumeurs etc… On ne peut répondre simultanément à toutes ces questions, par l'intermédiaire d'un test du moins : les tests ne permettent de répondre qu'à des questions simples.

2. ON NE CHOISIT PAS LE SEUIL DE SIGNIFICATION

   Que dirait-on d'un médecin annonçant : j'aime le risque alors j'ai choisi un risque

α de 0,4 et le traitement que je propose est efficace (ou actif) à ce risque ?
α = 0,05 est conventionnel

3. ON NE DIT PRATIQUEMENT JAMAIS : L'EXACTITUDE DE L'HYPOTHESE NULLE EST DEMONTREE

4. ON N'ENCHAINE PAS LES TESTS DE FAÇON INCONSIDEREE
En effet, les risques de conclusion à tort augmentent alors.
Par exemple, supposons que l'on veuille tester l'égalité à une valeur donnée de deux proportions (ex : succès d'une intervention chirurgicale dans deux services hospitaliers, le pourcentage de succès sur la France étant par ailleurs connu (données de l'année précédente par exemple)). Que se passe-t-il si l'on effectue deux tests successifs dont les hypothèses nulles sont :
service 1 : $\varphi_1 = \varphi_0$; puis service 2 : $\varphi_2 = \varphi_0$.
Le risque de première espèce global de la procédure exprime la probabilité de dire au moins une fois (soit au cours du premier test soit au cours du second) H_1 alors que H_0 est vraie les deux fois :
Pr(conclure H_1 au moins une fois si H_0 est vraie) = 1 - *Pr*(ne rejeter H_0 aucune des deux fois si H_0 est vraie)
Or *Pr*(ne pas rejeter H_0 si H_0 est vraie) = 1 - α
Donc *Pr*(ne rejeter H_0 aucune des deux fois si H_0 est vraie) = $(1 - \alpha)^2$
d'où un risque total = $1 - (1 - \alpha)^2$
Exemple
Si α = 0,05, le risque global est environ 0,10.
Cette situation s'aggrave si le nombre de tests s'accroît. Ainsi, dans le cas de
    - 3 services le risque global est 0,14
    - 10 services le risque global est 0,40
    - 100 services le risque global est 0,994

Cela signifie par exemple que dans le cas où 10 services sont à comparer à une référence il y a 4 chances sur 10 pour qu'au moins une fréquence observée s'écarte de façon significative de la valeur de référence, alors qu'en réalité tous les résultats sont homogènes. Si l'on prend la fréquence observée la plus différente de la valeur de référence, le test permettra de conclure, à tort, avec une probabilité supérieure à 0,4.
En fait, lorsque l'on désire faire des comparaisons multiples, des tests spécifiques doivent être utilisés de façon que les conclusions puissent être tirées avec un risque d'erreur α global de 5 %.

5. IL EST DANGEREUX ET ERRONE DE CHOISIR LES HYPOTHESES AU VU DES DONNEES

Lorsque l'on opère de cette façon, on a en réalité réalisé plus ou moins consciemment un nombre indéterminé de tests que l'on a jugés non concluants.

LA STRATEGIE D'ANALYSE DES DONNEES DOIT ETRE FIXEE CLAIREMENT AVANT LA REALISATION DE L'EXPERIENCE

Chapitre 12 - Quelques tests usuels

12.1 - Tests concernant des variables de Bernoulli

12.1.1 - Test d'égalité d'une proportion « vraie » à une valeur donnée (ou test de comparaison d'une proportion observée à une valeur donnée)

12.1.1.1 Mise en place du test

Exemple : les souris du chapitre précédent

1. Les hypothèses en présence

   H_0 (hypothèse nulle) : la proportion « vraie » (de souris cancéreuses dans la population des souris traitées) est égale à ϕ_0 (proportion hypothétique ou supposée qu'on se donne pour le test).
   H_1 (hypothèse alternative) : la proportion « vraie » est différente de ϕ_0.
   Notations :
   $H_0 : \phi = \phi_0$
   $H_1 : \phi \neq \phi_0$
2. Définition du paramètre

   $$Z = \frac{P_n - \varphi_0}{\sqrt{\frac{\varphi_0(1-\varphi_0)}{n}}}$$

   où P_n représente la variable aléatoire proportion.
   Sous H_0, Z est à peu près distribuée selon N(0, 1)

   [conditions de validité : $n\phi_0 \geq 5$ et $n(1 - \phi_0) \geq 5$]
3. Choix d'un seuil de signification α

   Construction de l'intervalle de pari de niveau $1 - \alpha$: $\text{IP}_{1-\alpha}$
   Exemple : $\alpha = 0{,}05$ $\text{IP}_{0,95}$= [-1,96 ; 1,96] (lu dans la table de la distribution normale)
4. Mise en place de la procédure de décision

   Lorsque les données seront disponibles on obtiendra une valeur du paramètre Z, soit :

   $$z = \frac{p - \varphi_0}{\sqrt{\frac{\varphi_0(1-\varphi_0)}{n}}}$$

Si $z \notin \mathrm{IP}_{1-\alpha}$ on rejette H_0 et on dit : au risque α l'hypothèse d'égalité de la proportion « vraie » et de la valeur donnée est fausse ; ou, au risque α, la proportion « vraie » est différente de la valeur donnée.
Si $z \in \mathrm{IP}_{1-\alpha}$ on ne rejette pas H_0 ou « on ne conclut pas ».

5. Recueil des données. Conclusion

Rappelons les conditions de validité : $n\phi_0 \geq 5$ et $n(1 - \phi_0) \geq 5$

12.1.1.2 Autre interprétation du paramètre z

Regardons la forme du paramètre *z*. On conclut (c'est-à-dire on rejette H_0) si $z \notin [-u_\alpha ; u_\alpha]$ c'est-à-dire si $|z| > u_\alpha$ soit si :

$$|p - \varphi_0| > u_\alpha \sqrt{\frac{\varphi_0(1-\varphi_0)}{n}}$$

c'est-à-dire si la proportion observée *p* est suffisamment différente de φ_0. Voilà pourquoi on dit que l'on compare *p* et φ_0. C'est pourquoi on dit aussi, lorsque H_0 est rejetée :

> La proportion observée est **significativement** différente de la valeur donnée, au risque α (0,05), ou encore : la différence entre *p* et ϕ_0 est significative. Ce qui indique une différence entre la valeur donnée et la proportion « vraie » ϕ.
> Lorsque H_0 n'est pas rejetée, on dit : la proportion observée n'est pas significativement différente de la valeur donnée.

Très important : une même différence $|p - \varphi_0|$ peut être ou non significative selon la valeur de *n*.

Si l'on vous demande : *p* = 0,25 et 0,2, sont-elles significativement différentes, ne répondez-pas ; demandez : quelle est la taille de l'échantillon sur lequel *p* a été calculé, à quel risque ?

12.1.1.3 Nombre de sujets nécessaires

Si on considère qu'en réalité $\varphi = \varphi_1$, le nombre de sujets nécessaires pour obtenir une puissance 1 - β (β < 0,5) est approximativement donné par

$$n = \frac{[1,96\sqrt{\varphi_0(1-\varphi_0)} + u_{2\beta}\sqrt{\varphi_1(1-\varphi_1)}]^2}{[\varphi_0 - \varphi_1]^2}$$

Conditions de validité : $n\varphi_0 \geq 5$ et $n(1 - \varphi_0) \geq 5$

Chapitre 13 - Tests concernant des variables qualitatives

Introduction

On a jusqu'à présent complètement négligé les variables qualitatives à plus de deux modalités. On a en effet toujours parlé de **moyenne**, et cette notion n'existe pas pour les variables qualitatives, sauf pour celles à deux modalités grâce à un artifice de codage. Il n'y a pas d'instrument permettant de résumer la distribution d'une variable qualitative ; il faut considérer la distribution dans son ensemble, c'est-à-dire l'ensemble des probabilités pour que telle ou telle modalité se réalise. Pourtant des problèmes de choix d'hypothèses se posent également dans le cas de telles variables ou tels caractères (ex : la répartition [distribution] de la couleur des cheveux diffère-t-elle chez les habitants de tel département et de tel autre ?). Si la répartition du caractère est connue dans une des deux populations, on aura à comparer une répartition « observée » à une répartition donnée. Si les deux répartitions sont inconnues, on aura à comparer deux répartitions « observées ». Ces problèmes sont respectivement les homologues des tests de comparaison d'une moyenne à une valeur donnée, de comparaison de deux moyennes. Il existe des tests adaptés à chacun de ces cas.

13.1 - Comparaison d'une répartition observée à une répartition donnée ou test du $\chi 2$ d'ajustement

13.1 - Comparaison d'une répartition observée à une répartition donnée ou test du $\chi 2$ d'ajustement

Supposons que l'on souhaite savoir si la répartition de la couleur des cheveux dans la population des habitants du département A diffère de la répartition de la couleur des cheveux dans la population française, cette dernière répartition étant supposée donnée. Supposons qu'il y ait k couleurs répertoriées. On est alors amené à considérer une variable qualitative à k modalités. Notons φ_i la probabilité de survenue de l'événement « la $i^{ème}$ modalité est observée ».

Exemple :

φ_1 = probabilité qu'un individu tiré au hasard dans le département A ait les cheveux blonds

φ_2 = probabilité qu'un individu tiré au hasard dans le département A ait les cheveux bruns

etc...

Notons par ailleurs φ_{hi} la proportion « vraie » de la modalité i dans la population française.

On s'apprête à réaliser une expérience sur n individus à l'issue de laquelle on disposera d'un ensemble de O_i (O_i = nombre d'individus présentant la modalité i du caractère étudié, parmi les individus de l'échantillon).

13.1.1 Les étapes de mise en œuvre

1. Les hypothèses en présence

Deux hypothèses sont en présence :

i. la répartition « vraie » de la variable dans la population étudiée coïncide avec la répartition donnée (hypothèse nulle H_0)

ii. les répartitions diffèrent (hypothèse alternative H_1)

Avec les notations précédemment introduites, cela s'écrit :

H_0 : hypothèse nulle : $\phi_i = \phi_{hi}$ pour tous les i de 1 à k.

H_1 : hypothèse alternative : $\phi_i \neq \phi_{hi}$ pour au moins une modalité, c'est-à-dire pour au moins un i.

2. Construction du paramètre

On a déjà mis en place ce test dans le cas d'une variable (0 - 1) c'est-à-dire d'une variable à deux modalités. Dans ce cas, les hypothèses en présence étaient bien du type ci-dessus c'est-à-dire
H_0 : $\phi = \phi_{h1}$ et $1 - \phi = \phi_{h2} = 1 - \phi_{h1}$

ce qui s'écrit avec les nouvelles notations :
$\phi_1 = \phi_{h1}$ et $\phi_2 = 1 - \phi_{h1}$

Mais on n'avait retenu que la condition $\phi = \phi_{h1}$ (en fait $\phi = \phi_0$) car dans ce cas les deux conditions ci-dessus sont redondantes.
Le paramètre calculé retenu était :

$$z = \frac{p - \varphi_{h1}}{\sqrt{\dfrac{\varphi_{h1}(1 - \varphi_{h1})}{n}}}$$

Calculons son carré

$$z^2 = \frac{n(p - \varphi_{h1})^2}{\varphi_{h1}(1 - \varphi_{h1})} = \frac{n(p - \varphi_{h1})^2}{\varphi_{h1}} + \frac{n(p - \varphi_{h1})^2}{1 - \varphi_{h1}}$$

$$z^2 = \frac{(np - n\varphi_{h1})^2}{n\varphi_{h1}} + \frac{(n(1-p) - n(1 - \varphi_{h1}))^2}{n(1 - \varphi_{h1})} = \frac{(np - n\varphi_{h1})^2}{n\varphi_{h1}} + \frac{(n(1-p) - n\varphi_{h2})^2}{n\varphi_{h2}}$$

Or np = nombre d'individus observés présentant la valeur 1 c'est-à-dire la modalité 1 de la variable ; or sous H_0la probabilité de cette modalité est ϕ_{h1}. On s'attend donc à observer $n\phi_{h1}$ individus présentant cette valeur. Ce nombre d'individus attendu s'appellera effectif attendu ou calculé de la première modalité et sera noté A_1.
De la même façon, $n(1 - p)$ = nombre d'individus observés présentant la valeur 0 c'est-à-dire la modalité 2 de la variable ; or sous H_0 la probabilité de cette modalité est $\phi_{h2} = 1 - \phi_{h1}$. On s'attend donc à observer $n\phi_{h2}$individus présentant cette valeur. Ce nombre d'individus attendu s'appellera effectif attendu ou calculé de la seconde modalité et sera noté A_2.

D'où $$z^2 = \frac{(O_1 - A_1)^2}{A_1} + \frac{(O_2 - A_2)^2}{A_2}$$

où les O_i représentent les effectifs observés dans les différentes modalités,
les A_i représentent les effectifs $n\phi_{hi}$dits prévus ou calculés ou **ATTENDUS** dans les différentes modalités.

GENERALISATION
Lorsque les variables considérées ont plus de deux modalités, on généralise le calcul ci-dessus et on retient le paramètre suivant :

$$Q = \sum_{i=1}^{k} \frac{(O_i - A_i)^2}{A_i}$$

où la somme s'étend à toutes les k modalités de la variable.
On rappelle que les O_i sont les effectifs observés, et que les A_i valent $n\phi_{hi}$.

On remarque que Q chiffre l'écart entre ce qui est prévu par l'hypothèse H_0 et ce qui est obtenu ; cet écart se fonde naturellement sur les différences $O_i - n\phi_{hi}$ car $n\phi_{hi}$ est le nombre attendu d'individus présentant la modalité i.

Exemple : si $\phi_{hi} = 0,4$, sur 100 individus on en attend 40 présentant la modalité i.
C'est le nombre que l'on aurait si la distribution d'échantillonnage coïncidait avec la distribution hypothétique.

Par ailleurs on a pu montrer (résultat dû à Pearson) que sous H_0 (et si tous les $A_i \geq 5$) ce paramètre a une distribution qui ne dépend que du nombre de modalités, k.
Cette distribution porte le nom de **DISTRIBUTION DE χ^2**.

Si bien que l'on peut former - grâce encore à une table - un intervalle de pari de niveau donné relatif à cette variable.

RETENONS :

CONDITIONS DE VALIDITE : TOUS LES A_i DOIVENT ETRE AU MOINS EGAUX A 5

3. Intervalle de pari

α étant choisi (0,05), construction de l'intervalle de pari $IP_{1-\alpha}$

La variable χ^2 a l'allure présentée figure 13. On remarque qu'il serait stupide de choisir l'intervalle de pari centré dessiné sur cette figure car alors des valeurs numériques voisines de zéro pour la valeur Q_c du paramètre$Q$ seraient dans la région critique du test ; or des valeurs proches de zéro sont plutôt compatibles avec H_0 d'où le choix suivant (voir figure 14) :

$$IP_{1-\alpha} = [0 ; K_{ddl,\alpha}]$$

C'est cette valeur, notée $K_{ddl,\alpha}$ qui est lisible directement dans une table.

Remarque : notez que cet intervalle, bien que non symétrique autour de la moyenne, respecte la définition d'un intervalle de pari donnée section 9.4.1.

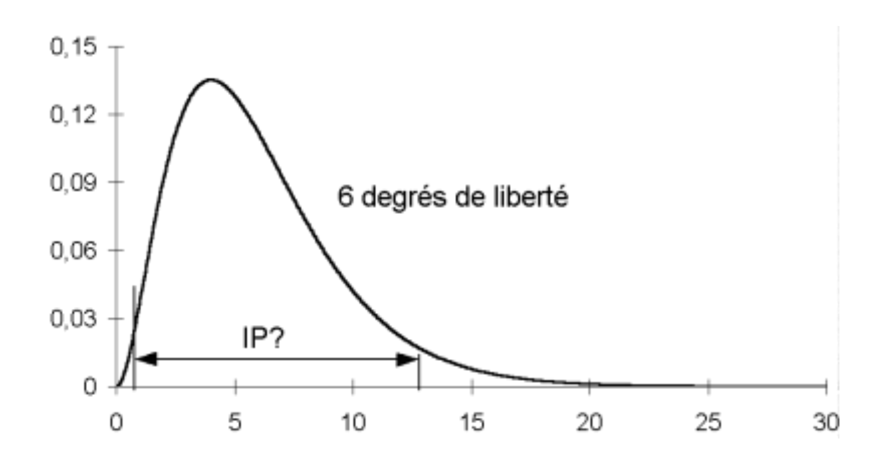


Figure 13 : distribution de χ^2

4.

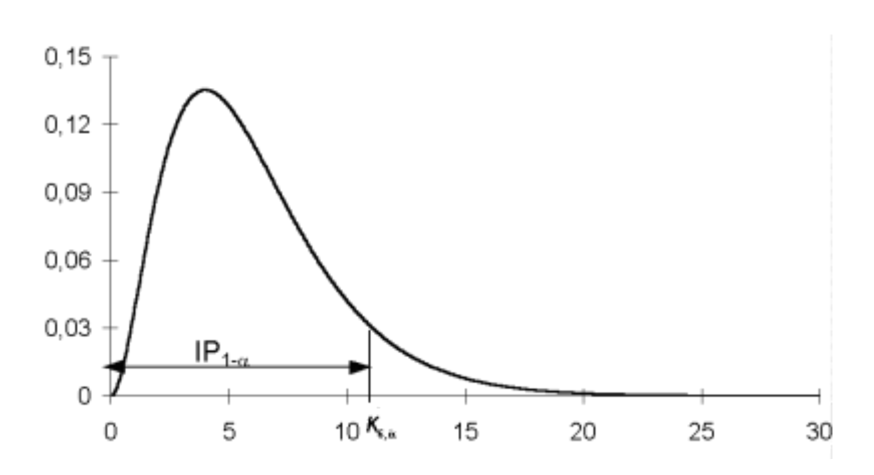


Figure 14 : distribution de χ^2

5. **Usage de la table**
   Cette table comporte - comme celle du *t* de Student - une entrée entière appelée nombre de degrés de liberté (ddl). On montre que pour le test envisagé ici

nombre de degrés de liberté = nombre de modalités - 1

   Exemple : $K_{5;0,05}$ (5 ddl, si 6 modalités) = 11,07
   La suite de la mise en place de ce test est usuelle.

6. Règle de décision

   Si $Q_c \leq K_{\text{ddl},\alpha}$ on ne conclut pas
   Si $Q_c > K_{\text{ddl},\alpha}$ H_0 est rejetée. Cela signifie que l'on conclut que la répartition du caractère étudié (par exemple la couleur des cheveux dans le département *A*) **ne coïncide pas** - ou **ne s'ajuste pas** - avec la répartition donnée (par exemple la répartition de la couleur des cheveux dans la population française). On admet, en formulant cette conclusion, un risque d'erreur égal à α.
7. Recueil des données et conclusion
   Exemple numérique : le tableau ci-dessous présente une application numérique de l'exemple considéré.

	couleur des cheveux			
	blonds	**bruns**	**roux**	**total**
effectifs observés (O_i)	25	9	3	37 (n)
effectifs attendus ($A_i = n\phi_{hi}$)	14,8	11,1	11,1	37
répartition donnée (ϕ_{hi})	0,4	0,3	0,3	1

8. Les conditions de validité sont vérifiées ($A_i \geq 5$).
   On obtient ici :

$$Q_c = \frac{(25-14{,}8)^2}{14{,}8} + \frac{(9-11{,}1)^2}{11{,}1} + \frac{(3-11{,}1)^2}{11{,}1} = 13{,}3$$

   On sait que *Q* est distribué selon un χ^2 à (3-1) degrés de liberté ; on lit dans la table : $K_{2;0,05}$ = 5,99.

Ainsi, la valeur calculée n'appartient pas à l'intervalle de pari : on conclut que la répartition du caractère ne coïncide pas avec la répartition donnée.

13.1 - Comparaison d'une répartition observée à une répartition donnée ou test du χ^2 d'ajustement

13.1.2 - Cas particulier : variable à deux modalités

On a vu que le paramètre du test Q généralise l'expression du carré du paramètre Z utilisé pour la comparaison d'une proportion observée à une valeur donnée. Dans le cas d'une variable à deux modalités (k = 2), ces deux paramètres sont égaux : $Q = Z^2$.

En outre, et sinon il y aurait incohérence, on peut vérifier l'égalité suivante :

$$K_{1;\alpha} = u_{\alpha}^{2}$$

Exemple : pour $\alpha = 0{,}05$ $K_{1;0,05} = 3{,}84 = (1{,}96)^2$
Ainsi, pour comparer une répartition observée à une répartition donnée, dans le cas d'une variable à deux modalités, on dispose de 2 tests équivalents, l'un fondé sur la distribution normale, l'autre fondé sur la distribution du χ^2 à 1 d.d.l. (qui est en fait la distribution du carré de N(0, 1)).
On peut utiliser l'un ou l'autre de ces tests indifféremment.
Exemple : Reprenons l'exemple du chapitre 11
Une race de souris présente un taux de cancers spontanés de 0,2. Sur 100 souris traitées on observe 34 cancers soit p = 0,34. La différence est elle significative ?

- test de comparaison :

$$z = \frac{0{,}34 - 0{,}2}{\sqrt{\frac{0{,}2 \times 0{,}8}{100}}} = 3{,}5$$

- test du χ^2 :

	cancer	absence de cancer	
répartition théorique	0,2	0,8	
effectifs attendus	20	80	
effectifs observés	34	66	100 (effectif total)

-

$$Q_c = \frac{(34-20)^2}{20} + \frac{(66-80)^2}{80} = 12{,}25 = (3{,}5)^2$$

Remarque : On parle souvent de ce test sous la terminologie « test du χ^2 d'ajustement » pour exprimer qu'il met à l'épreuve l'ajustement - la compatibilité - entre une répartition observée et une répartition donnée.

13.2 - Comparaison de plusieurs répartitions observées ou test du χ^2d'homogénéité

On reprend l'exemple précédent concernant la répartition de la couleur des cheveux mais sans plus supposer que l'une de ces répartitions est connue ; il s'agit par exemple des répartitions de ce caractère dans deux départements. On souhaite donc comparer deux répartitions observées. Pour cela, on s'apprête à réaliser une expérience mettant en jeu deux échantillons, un échantillon de n_1 individus issu de la population des habitants du département 1, et un échantillon de n_2 individus issu de la population des habitants du département 2. A l'issue de cette expérience on disposera d'un ensemble d'effectifs observés, notés de la façon suivante :

- O_{1i} est le nombre d'individus du premier échantillon présentant la modalité i de la variable.
- O_{2i} est le nombre d'individus du second échantillon présentant la modalité i de la variable.

Le test se met en place de la façon suivante :

1. Les hypothèses en présence

   H_0 : les répartitions « vraies » de la variable sont identiques dans les deux populations
   H_1 : les répartitions « vraies » sont différentes

   Ces hypothèses se schématisent par :

   H_0 : $\varphi_{1i} = \varphi_{2i}$ pour toutes les modalités i.
   H_1 : $\varphi_{1i} \neq \varphi_{2i}$ pour au moins une modalité i.
2. Construction du paramètre

   C'est encore ici le point délicat. La solution ressemble dans son approche à celle du problème de la comparaison de deux pourcentages. **Clé du principe** : on mélange les deux populations pour calculer une pseudo-répartition théorique. On se retrouve alors pratiquement dans la situation du paragraphe précédent. Cela se verra mieux sur un exemple. On va faire, pour des raisons de simplicité de calcul, une petite entorse à notre façon de procéder, et directement évaluer le paramètre dont on connaît la loi.
3. On construit ce que l'on appelle un **tableau de contingence** qui contient les résultats expérimentaux.
   On a procédé à une expérience portant sur 37 individus issus de la population 1 et 40 individus issus de la population 2. Les résultats sont les suivants :

Tableau 4 : effectifs observés (O_{1i} et O_{2i})

	blonds	bruns	roux	nombre total
échantillon 1	25	9	3	37 = n_1
échantillon 2	13	17	10	40 = n_2

4. On construit une pseudo-répartition de référence, en mélangeant les résultats expérimentaux, c'est-à-dire en oubliant leur origine (population 1 ou population 2). On obtient les résultats suivants, en termes d'effectifs (première ligne), puis en termes de fréquences (deuxième ligne).

Tableau 5 : répartition de « référence »

	blonds	bruns	roux	nombre total
mélange	38	26	13	77
fréquences	38/77 = 0,49	26/77 = 0,34	13/77 = 0,17	

5. Ces trois fréquences, 0,49, 0,34, 0,17, vont jouer maintenant le rôle des probabilités hypothétiques φ_{hi} de la section 13.1. Pour la commodité de l'écriture, on les note respectivement p_1, p_2, p_3.
6. On forme le tableau des effectifs attendus.
   Si l'hypothèse nulle est juste, c'est-à-dire si les répartitions de la couleur des cheveux coïncident dans les deux départements, on s'attend à trouver des effectifs calculés comme suit :
   effectif attendu pour la modalité i (modalité 1 = blond, modalité 2 = brun, modalité 3 = roux) dans l'échantillon j (j = 1 ou 2) : n_j multiplié par p_i
   Par exemple le nombre attendu d'individus bruns dans l'échantillon de la première population est : 37 × 0,34 = 12,6.

   En effectuant systématiquement ces calculs, on obtient le tableau des **EFFECTIFS ATTENDUS**.

Tableau 6 : effectifs attendus (A_{1i} et A_{2i})

	blonds	bruns	roux
échantillon 1	18,1 (n_1p_1)	12,6 (n_1p_2)	6,3 (n_1p_3)
échantillon 2	19,6 (n_2p_1)	13,6 (n_2p_2)	6,8 (n_2p_3)

7. On calcule finalement le paramètre du test
   On montre que le paramètre adapté à ce test est :

$$Q = \sum_{i=1}^{k} \frac{(O_{1i} - A_{1i})^2}{A_{1i}} + \sum_{i=1}^{k} \frac{(O_{2i} - A_{2i})^2}{A_{2i}}$$

où k demeure le nombre de modalités de la variable.

On a souvent recours à une expression plus compacte de l'expression ci-dessus et on écrit :

$$Q = \sum_{j=1}^{\text{nombre de cases du tableau}} \frac{(O_j - A_j)^2}{A_j}$$

8.

MAIS ICI LA SOMMATION S'ETEND A **TOUTES LES CASES** DES TABLEAUX, numérotées grâce à l'indice *j*.
Exemple : dans l'exemple traité il s'agira donc de calculer une somme de 6 termes.

On montre que, si H_0 est vraie, Q est distribué comme un χ^2 à (3 - 1) × (2 - 1) degrés de liberté [3 est le nombre de modalités, et 2 le nombre de répartitions]
La VALIDITE de ce résultat suppose que tous les **effectifs attendus A_j soient au moins égaux à 5**.

GENERALISATION

Les calculs ci-dessus se généralisent à un nombre quelconque de modalités *k*, à un nombre quelconque de populations *m*.
Le paramètre Q à calculer a alors la forme ci-dessus, où la somme comprend $k \times m$ termes.
La distribution de Q, sous H_0 est alors un **χ^2 à (k - 1)×(m - 1) degrés de liberté**.
Les conditions de validité du test sont : $A_j \geq 5,\ 1 \leq j \leq km$

9. La suite des étapes de mise en œuvre est classique.

La valeur observée de Q, notée Q_c, sera comparée à la valeur $K_{ddl;0,05}$:

- si $Q_c \leq K_{ddl;0,05}$ on ne conclut pas. Il n'est pas démontré que les deux répartitions « vraies » diffèrent.
- si $Q_c > K_{ddl;0,05}$ on conclut que les deux répartitions observées diffèrent significativement.

Suite de l'exemple : on obtient :

$$Q_c = \frac{(25-18{,}1)^2}{18{,}1} + \frac{(9-12{,}6)^2}{12{,}6} + \frac{(3-6{,}3)^2}{6{,}3} + \frac{(13-19{,}6)^2}{19{,}6} + \frac{(17-13{,}6)^2}{13{,}6} + \frac{(10-6{,}8)^2}{6{,}8}$$

soit : $Q_c = 9{,}96$

Or : $K_{2;0,05} = 5{,}99 \Rightarrow$ rejet de H_0. Les répartitions observées de la couleur des cheveux diffèrent significativement dans les deux populations.

Remarque 1 : Ce test s'appelle aussi test du χ^2 **d'homogénéité** de plusieurs

répartitions.

Remarque 2 : **Cas particulier de deux variables à deux modalités** : dans le cas où l'on considère deux variables à deux modalités, c'est-à-dire dans le cas où le tableau de contingence est à deux lignes et deux colonnes, on observe que le problème se réduit à un problème de comparaison de deux proportions observées. On montre que, dans ce cas, la valeur de Q coïncide avec le carré de la valeur de Z, Z étant le paramètre formé pour comparer directement ces proportions (voir chapitre 12).

13.3 - Test d'indépendance entre deux variables qualitatives

Reprenons l'exemple précédent et supposons que les populations 1 et 2, plutôt que de correspondre à des individus habitant le département 1 et le département 2, soient en fait :

- population 1 : population des individus ayant les yeux bleus
- population 2 : population des individus ayant les yeux verts

La question que l'on aurait résolue dans le paragraphe précédent aurait été :

la répartition de la couleur des cheveux diffère-t-elle dans les populations d'individus aux yeux bleus ou verts. Ou encore, la répartition de la couleur des cheveux diffère-t-elle selon la couleur des yeux ? Autrement dit : la variable couleur des cheveux dépend-elle statistiquement de la variable couleur des yeux ?

Maintenant supposons que l'on veuille répondre à cette question. Plutôt que de prendre un échantillon de la population des individus aux yeux bleus et un autre échantillon issu de la population des individus aux yeux verts, autant prendre un échantillon de la population générale (c'est-à-dire quelle que soit la couleur de ses yeux) et observer**conjointement** la couleur des cheveux et la couleur des yeux. Vues comme cela, les deux variables jouent bien des rôles symétriques et le problème est donc de mettre à l'épreuve leur indépendance.

1. Les hypothèses en présence.

   On formule naturellement deux hypothèses :

   Hypothèse H_0
   les deux variables étudiées (couleur des cheveux, couleur des yeux) sont indépendantes. Sous cette hypothèse, le fait d'avoir observé chez un individu la couleur de ses cheveux (respectivement la couleur de ses yeux) n'apporte aucune information sur la couleur de ses yeux (respectivement la couleur de ses cheveux).
   On pourra se reporter au chapitre 6 dans lequel ont été commentées ces notions d'indépendance.
   On notera que, comme dans tous les cas rencontrés jusqu'ici, cette hypothèse est une hypothèse fine qui engage un ensemble d'égalités.
   En effet, on sait que l'indépendance s'exprime par :
   Pr(la modalité de la couleur des cheveux est l et la modalité de la couleur des yeux est c) = *Pr*(la modalité de la couleur des cheveux est l) × *Pr*(la modalité de la couleur

des yeux est *c*), et ceci pour tous les choix possibles de *l* et *c*.

Remarque : on pourra vérifier que parmi les (nombre de modalités de la couleur des cheveux × nombre de modalités de la couleur des yeux) égalités qui en résultent, certaines sont redondantes, et que (nombre de modalités de la couleur des cheveux - l) × (nombre de modalités de la couleur des yeux - 1) égalités suffisent à exprimer les mêmes conditions.

Hypothèse H_1

les deux variables étudiées ne sont pas indépendantes.

Cette hypothèse exprime le contraire de H_0.

TRES IMPORTANT (des erreurs sont souvent commises)
HYPOTHESE NULLE : LES DEUX VARIABLES SONT INDEPENDANTES
HYPOTHESE ALTERNATIVE : LES DEUX VARIABLES SONT LIEES

2. Le paramètre du test

Le paramètre est encore *Q*, et s'exprime exactement comme précédemment, c'est-à-dire :

$$Q = \sum_{j=1}^{\text{nombre de cases du tableau}} \frac{(O_j - A_j)^2}{A_j}$$

Ici le nombre de cases du tableau de contingence est égal au produit du nombre de modalités de la première variable et du nombre de modalités de la seconde variable. Les effectifs attendus s'obtiennent exactement comme dans le cas du paragraphe précédent, ainsi qu'on peut le voir sur l'exemple numérique ci-dessous.

Un exemple numérique

Le tableau ci-dessous montre un exemple de tableau de contingence (D. Schwartz, *Méthodes statistiques à l'usage des médecins et des biologistes*, Flammarion (collection statistique en biologie et médecine), 3e édition, p79) ; cet exemple est similaire aux précédents, si ce n'est que l'on a considéré un plus grand nombre de modalités pour la variable couleur des cheveux, et que la nouvelle variable introduite (couleur des yeux) comporte trois modalités. Ces modalités remplacent les échantillons considérés dans la section 13.2. Ainsi, la modalité « bleu » par exemple peut être lue : « échantillon issu de la population des individus aux yeux bleus ». La taille de cet échantillon n'est cependant plus maîtrisée.

Couleur des yeux	Couleur des cheveux					fréquence
	blonds	**bruns**	**roux**	**noirs**	**total**	
bleus	25	9	7	3	**44**	**44/124**
gris	13	17	7	10	**47**	**47/124**
marrons	7	13	5	8	**33**	**33/124**
total	45	39	19	21	124	
fréquence	45/124	39/124	19/124	21/124	124/124	

3. Les effectifs attendus s'obtiennent comme précédemment. Ainsi, l'effectif attendu relatif au couple « blonds, marrons » sera : 45/124×33/124×124 = 11,9.

   REMARQUES

   i. Pour alléger les calculs, on peut remarquer que l'effectif attendu relatif à la cellule localisée ligne *l*, colonne *c* est égal au rapport
      - du produit du total de la ligne *l* et du total de la colonne *c*,
      - et du total général.

   ii. La somme des effectifs attendus, soit en ligne, soit en colonne, coïncide avec les mêmes sommes sur les effectifs observés. Cette remarque permet une vérification partielle des calculs.

   iii. Dans la présentation des calculs, on a procédé au « mélange » des résultats sans plus tenir compte de la couleur des yeux (ce qui conduit à sommer les lignes du tableau). On peut de façon équivalente mélanger les résultats expérimentaux sans plus tenir compte de la couleur des cheveux, ce qui conduira à sommer les colonnes du tableau de contingence pour obtenir la répartition de référence. On pourra vérifier que les résultats du calcul sont strictement les mêmes, ce que l'on attend compte tenu du rôle symétrique joué par les deux variables étudiées.

SOUS L'HYPOTHESE NULLE D'INDEPENDANCE entre les deux variables, *Q* EST DISTRIBUE SELON un χ^2 à :

(nombre de modalités de la première variable - 1) × (nombre de modalités de la seconde variable - 1)

DEGRES DE LIBERTE.

Les CONDITIONS DE VALIDITE sont encore : $A_j \geq 5$.

4. La suite des étapes est habituelle

   En particulier, la règle de décision s'établit comme suit :

   - si la valeur calculée de *Q*, notée Q_c, est inférieure à $K_{ddl,\alpha}$, on ne rejette pas l'hypothèse d'indépendance des deux variables.

- si la valeur calculée Q_c est supérieure à $K_{ddl,\alpha}$, on rejette l'hypothèse d'indépendance des deux variables. On dira alors que les deux variables sont liées, au risque α.

Exemple :

Dans l'exemple ci-dessus, la valeur de Q_c, résultant de la sommation de 12 termes, est 15,1.

Le nombre de degrés de liberté est : (4 - 1)×(3 - 1) = 6, la valeur de $K_{6;0,05}$ associée étant 12,6 (lue dans une table). On rejette donc ici l'hypothèse d'indépendance : couleur des cheveux et couleur des yeux sont liées, ou encore sont dépendantes.
Voyons une illustration de cette dépendance. Sur la base des données observées on a :
*P*r(yeux bleus) = 44/124 = 0,35
*P*r(yeux bleus / cheveux blonds) = 25/45 = 0,56
La connaissance de la couleur des cheveux (ici la modalité « blond ») modifie la répartition de la couleur des yeux (ici la fréquence de la modalité « bleu » qui évolue de 0,35 à 0,56). Le test indique que cette modification est significative. En réalité la valeur de Q_c ci-dessus chiffre dans leur ensemble les différences entre $Pr(A / B)$ et$Pr(A)$, c'est-à-dire les écarts de $Pr(A$ et $B)$ par rapport au produit $Pr(A)Pr(B)$, où A est un événement relatif à la couleur des yeux et B un événement relatif à la couleur des cheveux.

Chapitre 14 - Liaison entre deux variables continues : Notion de corrélation

14.1 - Introduction

Nous avons rappelé dans le chapitre précédent la notion fondamentale d'indépendance entre deux variables qualitatives et vu la façon dont cette indépendance pouvait être mise à l'épreuve lors d'une expérience. Dans le chapitre 12, les tests mis en œuvre faisaient intervenir une variable quantitative continue et une variable qualitative encore jugées dans leurs interdépendances. Il se trouve qu'il existe une autre classe de problèmes mettant en jeu encore deux variables aléatoires, mais cette fois-ci, deux variables continues. Considérons, par exemple, deux variables aléatoires, l'insuffisance rénale (avec deux valeurs ou modalités présence-absence) et l'insuffisance hépatique (avec les deux mêmes modalités). Supposons que l'on connaisse un indicateur de la fonction rénale (ou de certains de ses aspects), la clairance à la créatinine par exemple et un indicateur de la fonction hépatique (ou de certains de ses aspects) la bilirubinémie et que le diagnostic d'insuffisance rénale soit porté lorsque la clairance est inférieure à un seuil, celui d'insuffisance hépatique lorsque la bilirubinémie est supérieure à un autre seuil. On sait résoudre (voir chapitre 13) la question de savoir si les variables insuffisance rénale et insuffisance hépatique sont indépendantes ou liées. Toutefois, compte tenu des précisions données sur l'origine des diagnostics d'insuffisance rénale et d'insuffisance hépatique, on est tenté de reformuler le problème posé en ces termes : y a-t-il un lien entre les variables aléatoires *clairance à la créatinine* et *bilirubinémie* ? Un niveau élevé de l'une est-il « annonciateur » d'un niveau élevé de l'autre ? Ou encore : la connaissance du niveau de l'une modifie-t-elle l'idée que l'on se fait du niveau de l'autre, non encore observée ? Cette dernière formulation est très proche de la formulation utilisée pour discuter de l'indépendance entre événements : la connaissance du fait qu'un événement s'est réalisé (maintenant un niveau de clairance connu) modifie-t-elle la plausibilité d'un autre événement (maintenant la bilirubinémie) ?

Les situations dans lesquelles on se pose naturellement la question de savoir si deux variables continues sont liées sont extrêmement fréquentes. Voilà quelques exemples :

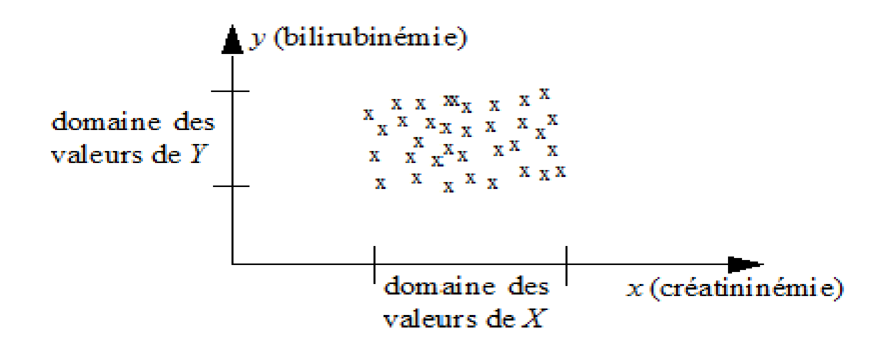
- la consommation de cigarettes (quotidienne ou cumulée) et la capacité respiratoire sont-elles liées ?
- la gastrinémie et la quantité de cellules ECL sont-elles liées ?
- les valeurs de glycémie obtenues selon deux méthodes de dosage sur les mêmes échantillons sanguins sont-elles liées [ici, il faut l'espérer].

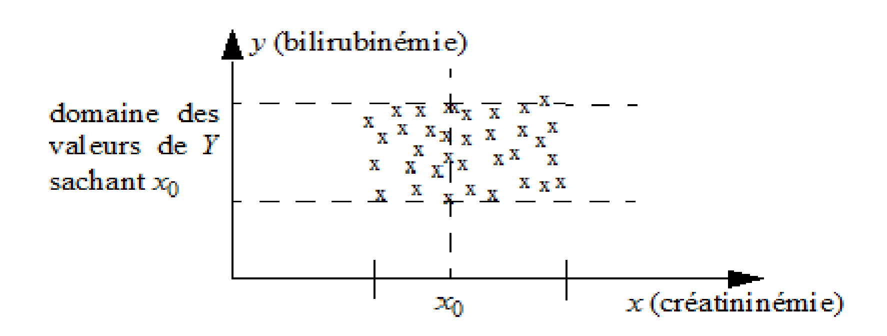
14.2 - Abord du problème

Considérons deux variables aléatoires continues X (créatininémie) et Y (bilirubinémie). Imaginons que nous ayons réalisé une expérience consistant en l'observation conjointe du niveau de ces deux variables sur un ensemble (échantillon) de n sujets. On dispose ainsi d'un ensemble de couples de valeurs x_i, y_i. La représentation naturelle - sinon la meilleure - de ces résultats est donnée dans la figure ci-dessous ; chaque couple de valeurs obtenu chez chaque individu est représenté par un point de coordonnées (créatininémie-bilirubinémie).

On lit sur un tel dessin, au moins grossièrement, le domaine des valeurs possibles de X, le domaine des valeurs possibles de Y.

Intéressons nous à un nouvel individu ; ne mesurons chez lui que la valeur de la créatininémie, x_0. Que peut-on dire alors, sur la base de cette connaissance et sur la base de l'expérience ci-dessus concernant le domaine des valeurs possibles de *Y* **pour ce même individu** ? On peut proposer la réponse géométrique ou visuelle indiquée sur la figure ci-dessous.

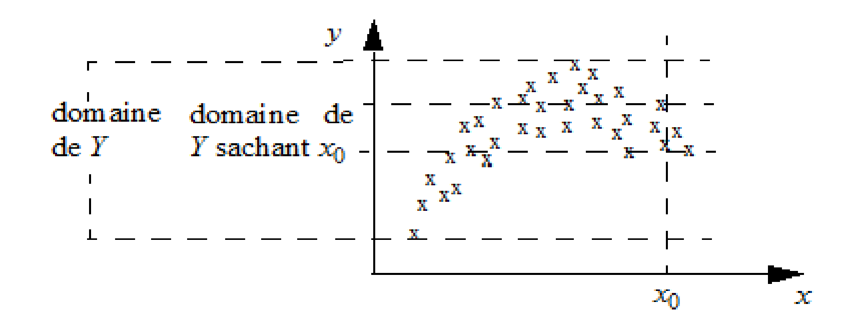


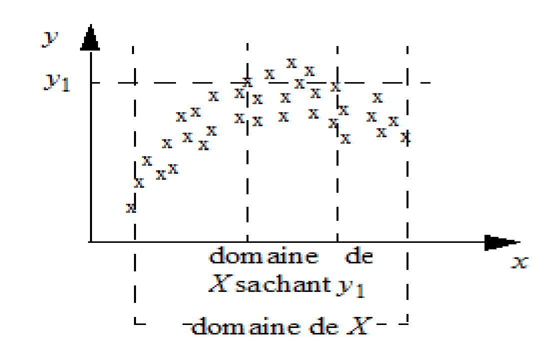


Le nouveau domaine possible - sachant x_0 - est très voisin du domaine initial ; ceci se reproduit pour toute valeur dex_0. Il est alors clair que dans cet exemple, la connaissance de *X* n'apporte pas d'information sur celle de *Y*. On a ici une situation visuelle d'un cas où les deux variables *X* et *Y* sont indépendantes. On pourrait renverser le rôle de *X* et *Y*, la conclusion serait la même.

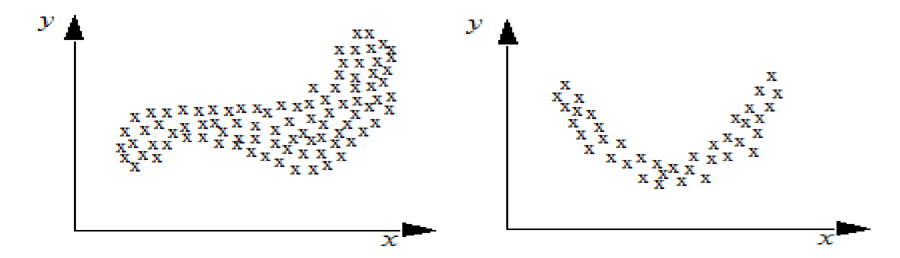
Considérons maintenant le cas où les résultats expérimentaux produisent la représentation de la figure ci-dessous.

Dans ce cas, au contraire, on voit clairement que la connaissance de x_0 (respectivement y_1) modifie le domaine des valeurs possibles, donc attendues de *Y* (respectivement *X*) ; les deux variables *X* et *Y* sont liées.

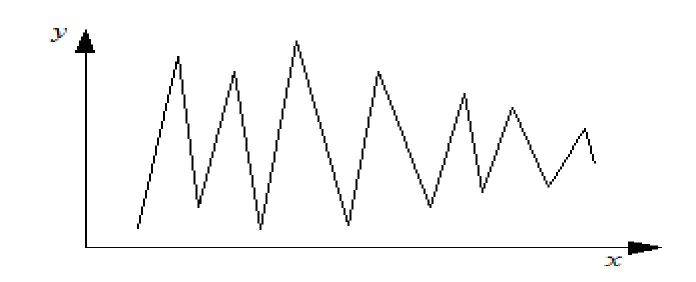




La modification ici concerne aussi bien l'amplitude du domaine que sa localisation en termes de valeurs.
L'appréciation visuelle de la dépendance correspond à l'appréciation de « l'épaisseur » de l'ensemble des points. Plus les points expérimentaux ont tendance à se répartir sur une courbe - non horizontale ni verticale - plutôt qu'à remplir une partie du plan, plus les variables sont liées.
Peut-on trouver un indicateur numérique de la force d'une telle liaison ? Au sens strict, la réponse est non.
Quelques situations de dépendance - c'est-à-dire de liaison - sont représentées sur les figures ci-dessous.



On ne sait pas, en toute généralité, résumer en un seul nombre exprimant la liaison entre deux variables continues les résultats d'une expérience.
On ne connaît qu'un indicateur général prenant en compte non pas le degré de proximité à une courbe quelconque mais le degré de proximité à une droite : c'est le coefficient de corrélation [linéaire].
Il faut voir cependant que dans la plupart des situations réelles au cours desquelles on s'intéresse à l'examen de la liaison entre deux variables, la possibilité d'interprétation des résultats est largement fonction du caractère monotone, sinon rectiligne, de la dépendance ; que dire en termes d'interprétation d'une dépendance figurée schématiquement sur la figure ci-dessous ?



14.3 - Un indicateur de covariation : le coefficient de corrélation

Cherchons alors à quantifier un phénomène de covariation, c'est-à-dire un phénomène de variation couplée entre X et$Y$.

On impose naturellement à l'indicateur recherché une invariance par translation : les phénomènes productifs de X et Ydemeurent fondamentalement inaltérés s'ils produisent $X + a$, $Y + b$. Ainsi l'indicateur se fondera-t-il sur les valeurs $x_i - m_x$ et $y_i - m_y$. Par ailleurs, on souhaite que l'indicateur ne dépende pas des unités exprimant X et Y ; alors on travaillera sur

$$x_{ri} = \frac{x_i - m_x}{s_X} \text{ et } y_{ri} = \frac{y_i - m_y}{s_Y}$$

Maintenant si Y et Y présentent un caractère de covariation, c'est que de façon fréquente, sinon systématique

- soit les variables varient dans le même sens, c'est-à-dire lorsque x_i est grand (i.e. x_{ri} positif par exemple), y_il'est également le plus souvent (i.e. y_{ri} positif), que lorsque x_i est petit ($x_{ri} < 0$) y_i l'est également ($y_{ri} < 0$) ; dans ce cas, le produit $x_{ri}y_{ri}$ est fréquemment positif.
- soit les variables varient en sens contraire : lorsque x_i est grand, y_i est petit, lorsque x_i est petit, y_i est grand ; dans ce cas le produit $x_{ri}\, y_{ri}$ est fréquemment négatif.

Compte tenu de l'analyse précédente, on choisit pour indicateur de la covariation ou corrélation le nombre :

$$r = \frac{1}{n-1}\sum_i x_{ri}y_{ri}$$

Ainsi

- si r est grand, c'est le signe d'une covariation dans le même sens de X et Y ;
- si r est petit (c'est-à-dire grand en valeur absolue et négatif), c'est le signe d'une covariation de X et Y en sens contraire ;
- si r est voisin de zéro, c'est le signe d'une absence de covariation.

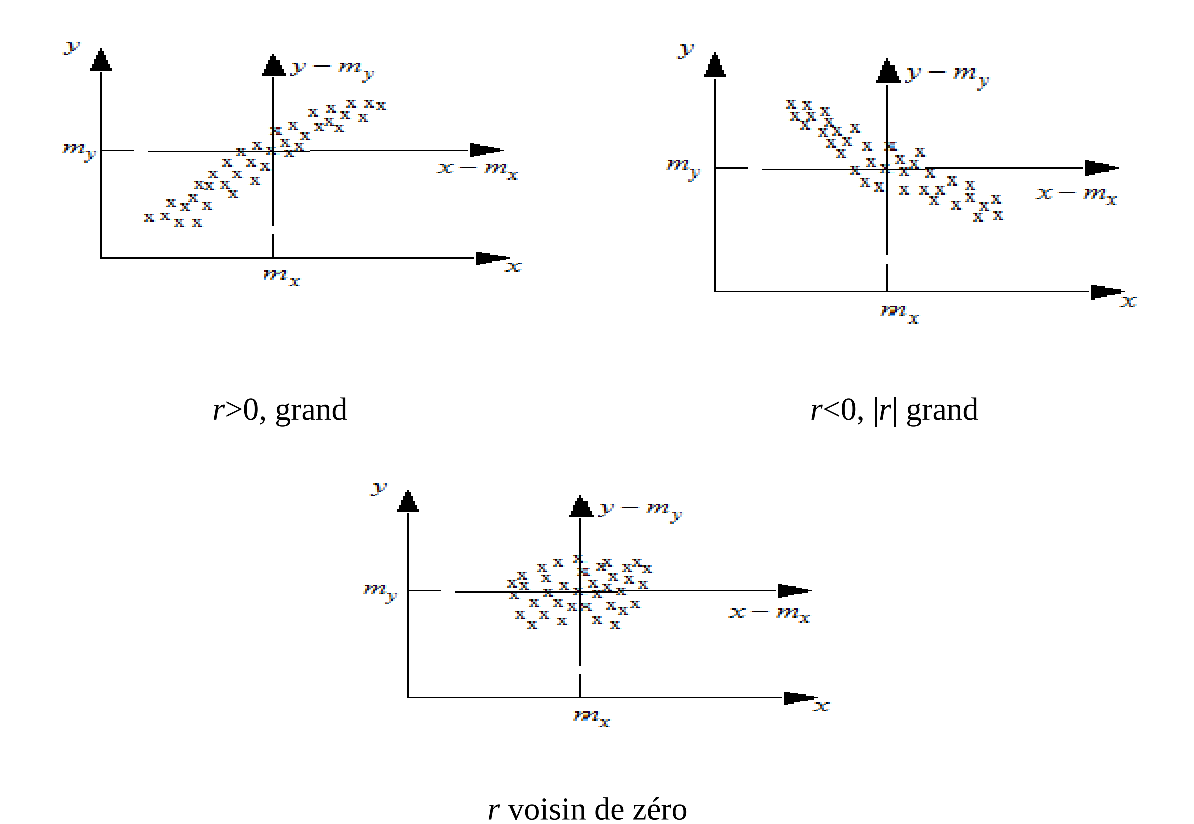
Retenons, exprimé sur la base des valeurs observées :

$$r = \frac{\frac{1}{n-1}\sum_i (x_i - m_x)(y_i - m_y)}{s_X s_Y}$$

Le numérateur de cette expression est appelé la covariance observée des deux variables X et Y, notée $cov_0(X, Y)$, dont on montre qu'elle s'exprime aussi sous la forme

$$cov_0(X, Y) = \frac{n}{n-1}\left(\frac{1}{n}\sum_i x_i y_i - \bar{x}\bar{y}\right)$$

Les figures ci-dessous présentent diverses situations relativement au coefficient de corrélation observé.



r>0, grand

r<0, |*r*| grand

r voisin de zéro

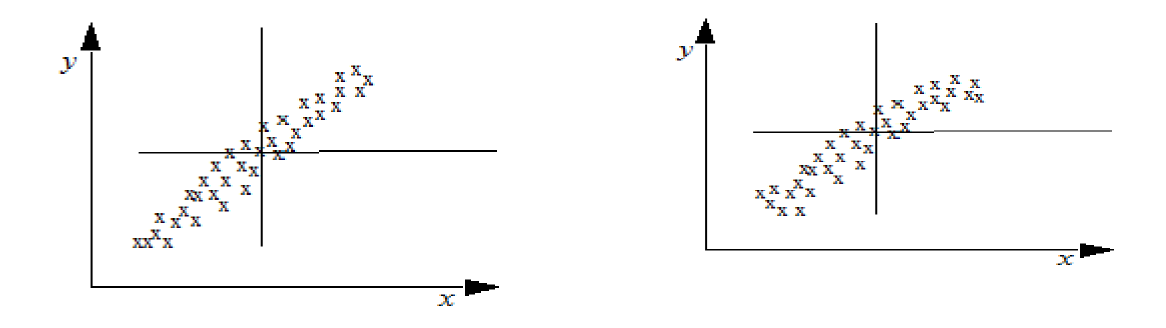
Propriétés numériques fondamentales de *r* :

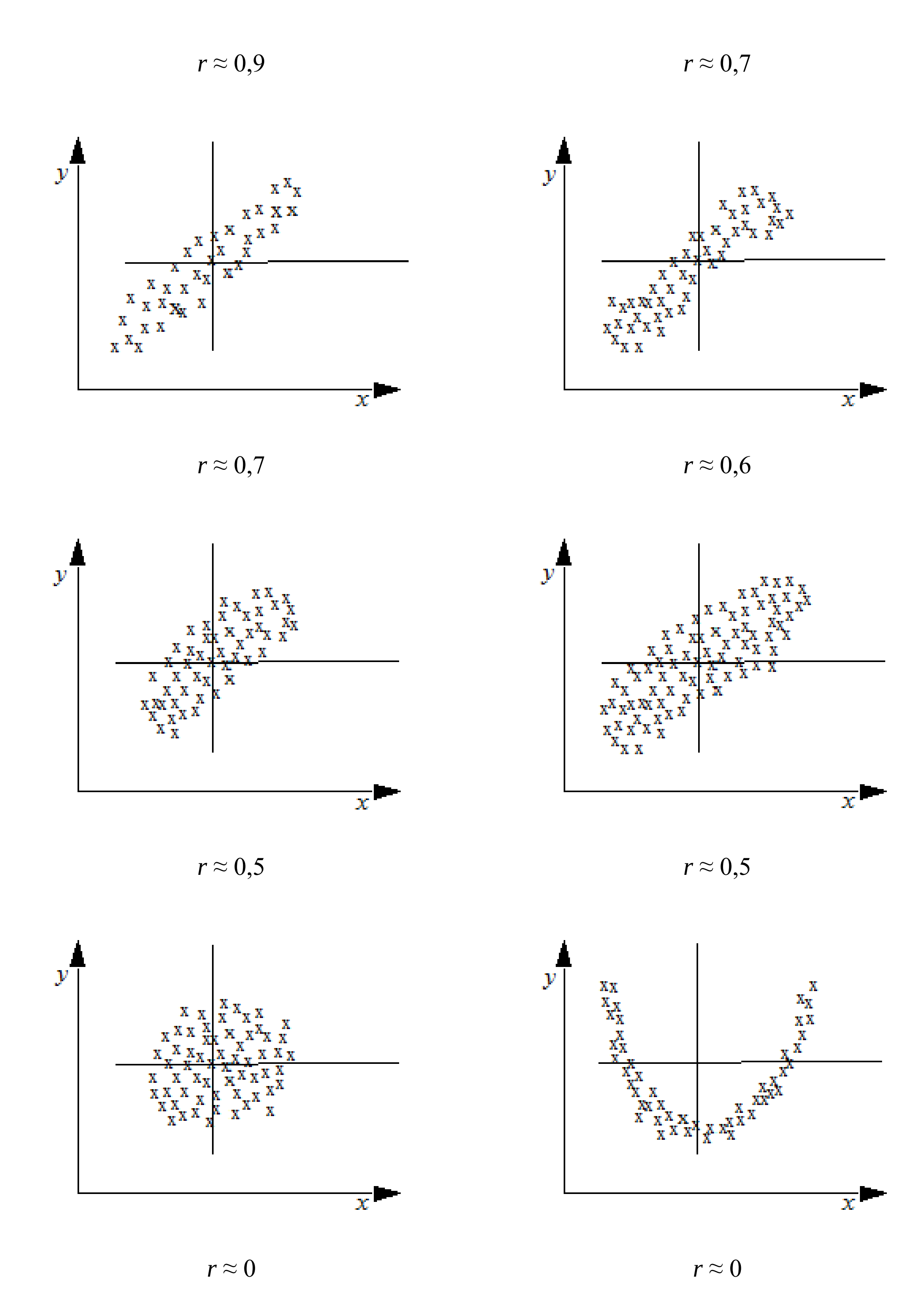
- *r* a toujours une valeur comprise entre -1 et 1 ;
- *r* prend la valeur -1 (respectivement 1) si et seulement si il existe des valeurs a et b telles qu'on ait pour tout $i y_i = a x_i + b$ avec a négatif (respectivement $a > 0$).

Remarques :

- plus *r* est grand en valeur absolue, plus les variables sont dites corrélées,
- la valeur absolue de *r* décroît,
  - lorsque s'estompe le caractère rectiligne du « nuage » des valeurs observées,
  - lorsque s'épaissit ledit nuage,
- une valeur absolue très faible du coefficient de corrélation ne permet pas de conclure à l'indépendance de deux variables. Deux variables indépendantes présenteront en revanche un coefficient de corrélation observé très faible en valeur absolue.

Quelques exemples sont présentés ci-dessous pour fixer les idées.





Remarque complémentaire :

Le coefficient de corrélation linéaire est, au même titre que toute statistique, soumis aux fluctuations d'échantillonnage. La question se pose alors de savoir que faire de cet indicateur en termes d'inférences. Par exemple, avant de conclure que les deux variables sont corrélées, peut-on se garantir du risque de l'observation d'un coefficient de corrélation nul sur une plus grande série d'observations ? On se retrouve dans le contexte des tests d'hypothèses avec ici une difficulté supplémentaire qui tient au fait que l'on n'a pas quitté le niveau expérimental,

le niveau intuitif. Il convient de trouver une contrepartie **« vraie »** à ce coefficient de corrélation observé *r*.

14.4 Le coefficient de corrélation « vrai »

Cherchons à substituer de la façon la plus naturelle possible des grandeurs « vraies » aux grandeurs observées constitutives de *r*. On note l'apparition au dénominateur de s_X et s_Y auxquelles on substitue naturellement σ_X et σ_Y, les écarts types « vrais » de *X* et *Y*. Au numérateur on remarque m_x et m_y auxquels on substitue $E(X)$ et $E(Y)$ les moyennes « vraies » de *X* et *Y*. Reste au numérateur une moyenne observée (lisons *n* à la place de *n*-1) ; on lui substitue une moyenne « vraie » : moyenne « vraie » du produit $[X - E(X)][Y - E(Y)]$, soit $E\{[X - E(X)][Y - E(Y)]\}$.

Cette moyenne « vraie » dépendant de *X* et *Y* à la fois s'appelle **covariance « vraie »** de *X* et *Y*.Finalement, on obtient la contrepartie « vraie » notée ρ :

$$\rho(X, Y) = \frac{E\{[X - E(X)][Y - E(Y)]\}}{\sigma_X \sigma_Y}$$

Remarque : à propos des notions d'espérance, de covariance « vraie », de coefficient de corrélation « vrai », voir le chapitre 6.

14.5 Test d'égalité du coefficient de corrélation « vrai » ρ à 0

Des calculs théoriques complexes, et imposant un certain nombre de restrictions, qui, dépassant le cadre de ce cours, ne seront pas mentionnés, permettent de calculer la distribution de *r* sous l'hypothèse - retenue comme hypothèse nulle - de nullité du coefficient de corrélation « vrai » ρ. Il s'agit d'une famille de distributions indexées par un entier appelé nombre de degrés de liberté. La mise en œuvre du test est alors conventionnelle :

- H_0 : $\rho = 0$ [les variables ne sont pas corrélées],
  H_1 : $\rho \neq 0$ [les variables sont corrélées]
- Paramètres du test : coefficient de corrélation observé
  $$r = \frac{\frac{1}{n-1}\sum_i (x_i - m_x)(y_i - m_y)}{s_X s_Y}$$
- sous H_0, *r* suit une distribution connue, dite du coefficient de corrélation à *n*-2 degrés de liberté où *n* est le nombre de couples (*xi*, *yi*) expérimentaux. L'intervalle de pari pour *r* est de la forme
  $IP_{1-\alpha} = [-\mathrm{corr}_\alpha(n-2) ; \mathrm{corr}_\alpha(n-2)]$, $\mathrm{corr}_\alpha(n-2)$ étant lue dans une table.
  Conditions de validité
  Les conditions de validité sont complexes et expriment que toute combinaison linéaire des variables *X* et *Y* est distribuée selon une loi normale. Autrement dit, toute variable $aX + bY$ où *a* et *b* sont deux nombres quelconques doit être normale.Pour la commodité de l'expression, on énoncera les conditions de validité sous le néologisme « distribution de (*X*, *Y*) binormale ».
- la suite de la mise en œuvre est standard.

Quelques exemples numériques

Au risque 5 % :

n = 10, $IP_{0,95}$ = [-0,632 ; 0,632], ddl = 8

n = 20, $IP_{0,95}$ = [-0,444 ; 0,444], ddl = 18

n = 50, $IP_{0,95}$ = [-0,280 ; 0,280], ddl = 48

Ainsi, par exemple, pour pouvoir conclure à la corrélation, lorsque l'on dispose de 20 observations (20 couples (x_i,y_i)), le coefficient de corrélation observé doit être supérieur à 0,444, ou inférieur à -0,444.

Autre formulation du test

On peut montrer que $t = r\sqrt{\frac{n-2}{1-r^2}}$ est, sous H_0, distribué selon une loi de Student à n-2 ddl.

Si on préfère utiliser ce paramètre plutôt que r, il faut lire la table de Student pour construire l'intervalle de pari.

Résumé du chapitre

1. La corrélation entre deux variables aléatoires quantitatives X et Y se mesure à l'aide du coefficient de corrélation « vrai » :
   $$\rho(X, Y) = \frac{E\{[X-E(X)][Y-E(Y)]\}}{\sigma_X \sigma_Y}$$
   Propriétés :
   - $\rho(X, Y) \in [-1 ; 1]$
   - Si X, Y indépendantes, alors $\rho(X, Y) = 0$
2. Disposant d'un échantillon de n couples (x_i, y_i) on définit le coefficient de corrélation observé :
   $$r = \frac{\frac{1}{n-1}\sum_i (x_i - m_x)(y_i - m_y)}{s_X s_Y} = \frac{\frac{n}{n-1}\left(\frac{1}{n}\sum_i x_i y_i - m_x m_y\right)}{s_X s_Y}$$
   Propriété : $r \in [-1 ; 1]$
3. Il existe un test de nullité du coefficient de corrélation « vrai » dont le paramètre est r.

4.Indépendance et corrélation sont des notions différentes ; deux variables dont le coefficient de corrélation « vrai » est nul peuvent être liées.

Chapitre 15 - Méthodologie des études épidémiologiques

15.1 La causalité

La causalité est une thématique centrale en philosophie des sciences et en logique, et les premiers écrits sur ce sujet remontent à Aristote. Jusqu'au 18ème siècle, la causalité nait de l'observation, et les connaissances sont construites à partir des observations sans idée préconçues du réel (inférence dite inductive). Nous formons alors une sorte d'anticipation, qui nous représente que le second événement (l'effet) doit se produire quand le premier (la cause) se produit - même si les mécanismes explicatifs liant ces deux événements nous échappent. Dans la vie courante, l'acquisition de nos apprentissages ou l'application des règles de « bon sens », illustrent cette conception de la causalité. Hume, philosophe écossais, montrera pourtant les limites de ce principe et l'impossibilité de prouver la relation causale de l'observation de la succession de deux événements dont on ne peut jamais exclure la coïncidence.

Le questionnement sur la causalité en médecine est également ancien. Claude Bernard écrit en 1865 :

« *L'esprit de l'homme ne peut concevoir un effet sans cause, de telle sorte que la vue d'un phénomène éveille toujours en lui une idée de causalité. Toute la connaissance humaine se borne à remonter des effets observés à leur cause. À la suite d'une observation, une idée relative à la cause du phénomène observé se présente à l'esprit ; puis on introduit cette idée anticipée dans un raisonnement en vertu duquel on fait des expériences pour la contrôler.* »

On retrouve dans ce propos une conception différente de la causalité : il s'agit de déduire (au sens strict) les conséquences d'une hypothèse et ensuite de comparer ces conséquences aux données. S'il y a désaccord, alors l'hypothèse est réfutée. Dans le cas contraire, l'hypothèse n'est pas prouvée mais notre croyance en elle s'en trouve renforcée. Cette conception, de type déductif, formalisée par Karl Popper au début du 20ème siècle, s'est imposée comme le socle de la découverte scientifique.

Des livres d'épidémiologie entiers, dont certains très mathématiques utilisant les outils de la logique, traitent de ce problème de la causalité. On retiendra qu'une relation causale entre deux caractères pourra être évoquée lorsque l'un des deux est « contrôlé ». L'essai contrôlé est la seule méthode qui permet de mesurer l'effet causal d'une intervention, par exemple un traitement, sur un événement, par exemple, la guérison d'une maladie.

15.2 Démarche expérimentale et démarche d'observation

Caractère contrôlé ; caractère aléatoire

Dans ce qui suit, les termes caractère, caractéristique, variable et facteur sont considérés comme synonymes.

On dit d'un **caractère** qu'il est **contrôlé** lorsque sa détermination nous appartient. **Exemple** : on s'intéresse à l'effet d'un traitement sur la survenue d'un type de cancer chez des souris. Le caractère absence ou présence du traitement est

contrôlé car déterminé par l'expérimentateur.
Dans le cas contraire, on dit que le **caractère** est **aléatoire**. **Exemple** : la survenue du cancer chez la souris.
Lorsqu'on envisage un problème de liaison entre deux variables (cela recouvre tous les problèmes que l'on a rencontrés) un au plus des caractères peut être contrôlé.

Démarche expérimentale

Lorsque l'expérience se conduit avec un facteur contrôlé, on dit que l'on suit une **démarche expérimentale**. Dans ce cas, au cours de la constitution de l'échantillon qui permettra de mettre en œuvre les tests, on décide du choix de la valeur d'un caractère (par exemple, on décide si le Xème patient sera traité ou non, et on étudie la guérison de la maladie).

Démarche d'observation

Lorsque l'expérience se conduit sur la base de deux facteurs aléatoires, on dit que l'on suit une **démarche d'observation** (par exemple, on observe si le Xème sujet est fumeur ou non, et on étudie la survenue de cancer).

Principe fondamental

La discussion de la causalité ne se conçoit pas sans contrôle d'un des deux caractères étudiés.
Autrement dit, on ne peut mesurer un effet causal hors d'une démarche expérimentale.
Seule cette démarche, en effet, permet d'assurer que les individus constituant l'échantillon sont comparables en tout (homogènes) sauf pour ce qui concerne le caractère contrôlé. Encore faut-il assurer cette homogénéité et la méthode de référence est le **tirage au sort**. On parle de **randomisation** pour l'attribution par tirage au sort du caractère contrôlé, le traitement.

15.3 - Les essais randomisés

15.3.1 Définition

Le but de l'essai randomisé est, grâce à une démarche expérimentale rigoureuse, d'évaluer l'efficacité d'une intervention de santé, par exemple un nouveau traitement - on parle alors d'essai thérapeutique randomisé. Il peut s'agir également d'une autre intervention médicale, par exemple une technique chirurgicale, un programme d'éducation pour la santé, un dispositif médical, une méthode diagnostique.

Dans un essai thérapeutique pour une maladie, on cherche le plus souvent à montrer qu'un nouveau traitement a une efficacité supérieure à celle du traitement habituellement utilisé dans cette maladie, ou à l'absence de traitement (notamment lorsqu'il n'existe pas encore de traitement d'usage pour la maladie étudiée). Parce que l'attribution du traitement est contrôlée et décidée par tirage au sort, on sera le cas échéant en mesure de conclure qu'une différence d'efficacité est causée par le traitement (au risque d'erreur statistique choisi). La réalisation d'un essai thérapeutique nécessite donc de définir le critère que l'on utilisera pour juger de l'efficacité - appelé **critère de jugement**. Des exemples de critères de jugement sont la régression des symptômes d'une maladie, la durée de survie, la valeur d'une constante biologique, un score de qualité de vie .... En pratique on distingue les **critères de jugement objectifs**, reposant sur une grandeur évaluable par une mesure physique objective, par exemple la charge virale, la concentration de cholestérol sanguin, la survie, ...

des **critères de jugement subjectifs**, qui peuvent être influencés par des effets de type psychologique, par exemple la douleur, l'anxiété, la qualité de vie, la mesure d'une impotence à une épreuve de marche, ... L'essai thérapeutique doit être organisé afin de limiter au maximum les biais pouvant modifier le déroulement de l'essai ou l'interprétation du critère de jugement.

15.3.2 Comment limiter les biais dans le déroulement d'un essai thérapeutique randomisé ? Aveugle et placebo

Le **placebo** est une présentation à l'identique du médicament à l'étude mais qui ne contient pas le principe actif. Ainsi lorsqu'un essai randomisé contre placebo est réalisé, un sujet dans l'essai recevra au hasard soit le médicament à l'étude soit une copie de ce médicament, et ni ce sujet, ni le médecin qui le prend en charge dans l'étude, ni la personne (souvent le médecin) qui évaluera le critère de jugement ne sauront ce qui est réellement pris. On parle dans ce cas de **mise en insu ou d'aveugle**. Les sujets recevant le placebo constituent ce que l'on appelle le groupe **témoin**(ou groupe contrôle par anglicisme). L'aveugle est justifié pour deux raisons principales : 1) d'une part, parce qu'il garantit que les patients seront a priori pris en charge et suivis de la même manière dans l'essai quel que soit le groupe dans lequel ils ont été randomisés. En l'absence d'aveugle, un patient recevant, par exemple, un placebo, pourrait être amené plus facilement à interrompre rapidement sa participation dans l'essai (puisqu'il a le placebo) pour pouvoir bénéficier d'un autre traitement, contrairement à un patient recevant le traitement à l'étude ; un patient recevant le traitement à l'étude pourrait faire l'objet de plus d'attention que celui du groupe placebo et 2) d'autre part, parce qu'il permet d'éviter une interprétation tendancieuse ou biaisée du critère de jugement si celui-ci est subjectif. L'évaluateur d'une douleur résiduelle 6 mois après le début de l'essai sera inévitablement influencé dans son jugement s'il connaît le groupe du patient, et aura tendance à trouver une plus grande efficacité sur la douleur chez les patients qui reçoivent le traitement à l'étude que chez les patients du groupe témoin.

D'ailleurs, en cas d'absence d'aveugle, le simple fait de soumettre un sujet à un traitement améliore souvent un critère de jugement subjectif, même si ce traitement n'a aucune efficacité intrinsèque : c'est **l'effet placebo**.

L'effet placebo est défini comme l'écart positif constaté entre l'effet thérapeutique observé et l'effet pharmacologique propre d'un médicament. Si l'on donne une substance inerte sur le plan pharmacologique, on observera uniquement un effet placebo. On considère ainsi que l'homéopathie ou d'autres médecines « douces », relèvent uniquement de l'effet placebo et donc que l'effet de ces thérapeutiques est exclusivement subjectif. Dans un essai en aveugle, l'effet placebo est réparti de la même manière entre les deux groupes de patients, et la différence observée est donc imputable aux seules propriétés pharmacologiques du traitement.

Un essai contre placebo en aveugle peut porter sur d'autres interventions que le seul médicament. Un exemple célèbre récent porte sur la chirurgie du genou chez les sujets obèses. Jusque récemment, tous les essais indiquaient une amélioration franche de la douleur et de la mobilité chez les sujets randomisés dans le groupe ayant subi l'intervention chirurgicale ligamentaire par rapport à ceux randomisés dans le groupe sans intervention.

Jusqu’au jour où un essai fut conduit comparant des sujets chez qui l’intervention sur les ligaments était réalisée, à des sujets endormis et chez lesquels un simulacre d’intervention (ouverture/fermeture simple de la cavité articulaire) était réalisé. Le résultat montrait l’amélioration de la douleur dans les deux groupes (l’effet placebo), et l’absence de différence entre les deux groupes (donc l’absence d’efficacité intrinsèque de l’intervention).

15.3.3 Comment limiter les biais dans l’analyse d’un essai thérapeutique randomisé ? Intention de traiter

L’analyse des résultats d’un essai thérapeutique est effectuée grâce à la méthode statistique. Ce sont les tests statistiques qui permettent de conclure si les différences observées entre le groupe de sujets traités et le groupe de sujets non traités ou recevant le placebo permettent de rejeter ou non l’hypothèse d’égalité de traitement, c’est-à-dire de conclure qu’un traitement est ou non meilleur que l’autre. De façon tout à fait générale, les tests employés sont des tests bilatéraux. De façon tout aussi générale, dans ce type d’essai de supériorité ou d’efficacité, l’analyse statistique est effectuée selon le **principe de l’intention de traiter**. Ce principe consiste à considérer le patient dans l’essai littéralement « tel qu’on avait l’intention de le traiter », c’est à dire, dans le groupe dans lequel il avait été randomisé - peu importe ce qui a été réellement pris. Par exemple, un patient randomisé dans le groupe placebo pourrait en réalité recevoir le traitement à l’essai pendant toute la durée de l’étude, par simple erreur au moment de la délivrance du médicament : quand bien même, ce patient sera analysé dans le groupe placebo. Le principe de l’analyse en intention de traiter a pour objectif de préserver les bénéfices de la randomisation au moment de l’analyse des résultats, et de limiter les biais liés aux écarts entre le protocole de l’essai et la conduite de celui-ci. En pratique, la conséquence implicite d’une analyse en intention de traiter sera que tout patient randomisé sera inclus dans l’analyse y compris ceux n’ayant pas bien voire pas du tout pris le traitement qui leur était alloué.

Cependant, les études randomisées ne sont pas toujours réalisables d’où l’importance de la question de causalité dans les études observationnelles en médecine. En particulier, si l’expérimentation peut parfois permettre de mesurer un effet causal entre un traitement et une maladie, elle est très souvent impossible lorsque la cause potentielle étudiée est un risque.

15.4 - Les études d’observation

On regroupe sous cette terminologie, les études dont l’objectif est d’identifier les facteurs associés à des événements de santé et ne reposant pas sur une démarche expérimentale. Il est en effet contre les principes éthiques d’exposer une personne à une cause potentielle de maladie pour étudier comment celle-ci survient - par exemple faire fumer par tirage au sort des sujets pour étudier le rôle du tabac dans les cancers. En revanche, les personnes d’elles-mêmes se répartissent souvent en exposées / non exposées (ex : fumeur/non fumeur), et le but de ces études d’observation sera donc d’étayer avec la meilleure démarche possible l’hypothèse testée à partir de ces observations.

Les études d’observation peuvent être catégorisées en cohortes - cas-témoins et transversales ; prospectives et rétrospectives. On notera que cette terminologie peut

s’appliquer également à la démarche expérimentale : un essai thérapeutique est une étude prospective imposant un recueil de données longitudinales. Dans les études d’observation, on étudie le plus souvent l’association entre une exposition et un événement de santé.

L’exposition peut être un risque (par exemple fumer, un polymorphisme génétique), ou un bénéfice (par exemple un traitement, le poids). Des événements de santé sont par exemple, la survenue d’une maladie, une guérison, un décès, une récidive, etc ...

15.4.1 Cohortes - Cas-témoins et études transversales

Etudes de cohorte

Une cohorte était le dixième d’une légion romaine. C’est plus généralement un ensemble de sujets. **Dans une étude dite de cohorte les sujets sont répartis en groupes en fonction de leur exposition** (par exemple, fumeur/non fumeur) et l’événement n’est pas survenu au moment où cette répartition est faite. Chacun de ces groupes définit une « sous cohorte », et la comparaison du taux de survenue de l’événement entre ces différentes sous cohortes, permettra de mesurer l’association entre exposition et événement. D’un point de vue pratique l’étude de cohorte est la démarche d’observation la plus « proche » de l’essai randomisé, la principale différence étant que dans un essai, l’attribution de l’exposition (le traitement) est réalisée par tirage au sort.

Etudes cas-témoins

Dans une étude cas-témoins (ou cas-contrôle), les groupes de sujets sont constitués en fonction de leur réalisation ou non de l’événement de santé : les cas sont par exemple les malades atteints d’un cancer et les témoins, des sujets non atteints de ce cancer. On compare les niveaux d’exposition dans ces deux groupes pour étudier l’association entre exposition et événement de santé. En général, on choisit de un à 4 témoins pour chaque cas et la proportion de malades dans l’étude est complètement déterminée (de 50 % pour 1 témoin pour 1 cas, à 20 % pour 4 témoins par cas), et ne correspond en rien à la proportion de malades dans la population cible.

Etudes transversales

Une étude transversale est une étude descriptive dont le principe est essentiellement de recueillir simultanément des informations sur expositions et événements de santé sur un échantillon représentatif de la population cible - celle à laquelle on souhaite pouvoir extrapoler les résultats. Les enquêtes de prévalence sont un exemple typique de ces études transversales, dans lesquelles on évalue le nombre de malades présents à un instant dans la population, et qui identifie les facteurs associés aux variations de prévalence. Ces études transversales sont

limitées par l'absence de description temporelle des expositions (et des événements), mais peuvent permettre d'identifier des relations entre événement de santé et exposition lorsque celles-ci sont invariables dans le temps (par exemple, le sexe, le groupe sanguin, ...).

15.4.2 Etudes prospectives et rétrospectives

Une étude est dite **prospective** lorsque l'exposition est mesurée avant la survenue de l'événement étudié. Une étude est dite **rétrospective** lorsque la mesure de l'exposition survient après la survenue de l'événement. Par exemple, une étude dans laquelle on mesure la consommation journalière de tabac, et dans laquelle on observe la survenue de cancer au cours du suivi des sujets est une étude prospective ; a contrario, demander la consommation de tabac des dix dernières années à des sujets ayant un cancer est une étude rétrospective. On notera que cette définition de prospective/rétrospective n'est pas consensuelle, et les épidémiologistes modernes recommandent de ce fait, de ne plus utiliser cette terminologie.

15.4.3 Données longitudinales

On dit que les données sont longitudinales lorsque qu'il existe plusieurs mesures à travers le temps par sujet. Par exemple, la mesure du taux de cholestérol chez un sujet tous les 6 mois, ou la mesure du statut fumeur/non fumeur au cours du temps est une donnée longitudinale. Les mesures longitudinales chez un même sujet ne peuvent pas être considérées comme réalisation de variables aléatoires indépendantes ; par exemple, le taux de cholestérol d'un sujet à un instant quelconque apporte une information sur le taux de cholestérol du même sujet 6 mois plus tard. Il faudra tenir compte de cette non-indépendance dans les analyses statistiques de ces données - qui seront souvent complexes.

15.4.4 En pratique

Le plus souvent, une étude de cohorte sera prospective, et aura recueilli des données longitudinales. Le plus souvent une étude cas-témoins sera rétrospective.

15.5 - Mesures d'association utilisées en épidémiologie

On traite le cas le plus simple où une exposition est répartie en deux niveaux (oui/non, présent/absent, exposé/ non exposé), et on notera E+ l'exposition, E- l'absence d'exposition au facteur étudié.

L'événement d'intérêt est également catégorisé en deux niveaux, M+ pour malade, M- pour non-malade. On notera que dans le cas d'un essai thérapeutique E+ est le traitement à l'étude, et M- peut être défini comme le succès thérapeutique, donc ce qui suit s'applique aussi bien à l'essai randomisé qu'aux études d'observation.

A partir de cette catégorisation, il est possible de dresser le tableau suivant :

	M+	**M-**
E+	n1	n2
E-	n3	n4

On définit

- le **risque absolu** chez les exposés, comme la proportion vraie de malades parmi les exposés
  P(M+ | E+), estimé par n1/(n1+n2)
- le risque absolu chez les non exposés, comme la proportion de malades chez les non exposés, P(M+ | E-), estimé par n3/(n3+n4)
- le **risque relatif** est une mesure d'association, défini comme le rapport des risques absolus chez les exposés et non exposés, P(M+ | E+) / P(M+ | E-).
  Ce risque est estimé par n1/(n1+n2) / n3/(n3+n4)
- le **rapport des cotes** (odds-ratio en anglais) est une autre mesure d'association très utilisée en biomédecine. Rappelons que la cote d'un événement est définie comme le rapport de sa probabilité sur son complémentaire : jouer à une cote de 9 contre 1 signifie jouer avec 9 chance de perdre contre une chance de gagner.
  Le rapport des cotes est défini comme le rapport de la cote de la maladie chez les exposés P(M+ | E+)/P(M- | E+) sur la cote de la maladie chez les non-exposés P(M+ | E-)/P(M- | E-), mais aussi, par application du théorème de Bayes, comme le rapport de la cote des expositions chez les malades P(E+ | M+)/P(E- | M+), par la cote des expositions chez les non malades P(E+ | M-)/P(E- | M-). Il est estimé par le rapport des produits croisés (n1n4) / (n2n3).

Le rapport des cotes est la seule quantité pertinente qui peut être estimée dans une étude cas-témoins puisque le nombre total de sujets non malades est déterminé par le nombre de témoins choisi par cas. Si la maladie est rare dans la population cible, aussi bien chez les exposés que chez les non-exposés, P(M+) est proche de 0 et donc P(M-) voisin de 1, et P(M+ | E+)/P(M- | E+) est voisin de P(M+ | E+) ; P(M+ | E-)/P(M- | E-) proche de P(M+ | E-) et donc le rapport des cotes défini ci-dessus est proche de du risque relatif.

Le risque relatif et le rapport des cotes sont des quantités qui peuvent prendre les valeurs entre 0 et l'infini. Sous l'hypothèse (nulle) d'indépendance entre l'exposition et l'événement étudié, ces deux quantités valent 1.

Un risque relatif ou un rapport de cotes supérieur à 1 (conclusion que l'on portera après avoir fait le test d'hypothèse approprié) signifie que l'exposition est un facteur de risque de l'événement étudié. Un risque relatif ou un rapport de cotes inférieur à 1 signifie que l'exposition est un facteur protecteur de l'événement. Un risque relatif de 50 (par exemple) pour l'exposition « fumeur » et l'événement « cancer du poumon » s'interprète littéralement comme « il y a 50 fois plus de cancer du poumon chez les fumeurs que chez les non fumeurs ».

15.6 - Risque attribuable, proportion de cas évitables

Le risque attribuable à un facteur est la proportion des cas que l'on pourrait éviter en supprimant ce facteur, lorsqu'il est causal. Par exemple, le risque de cancers du poumon attribuable au tabac est de l'ordre de 90 % (si on supprimait par magie le tabac, environ 10 % des cancers des bronches surviendraient encore, car ils sont dus à d'autres facteurs).

Soit :

- N le nombre total de malades ;
- N_F le nombre de malades exposés au facteur F ;
- N_{NF} le nombre de malades non exposés au facteur F.

On peut écrire $N = N_F + N_{NF}$.

Parmi les $N\ Pr(M)$ cas totaux de maladie, on ne peut éviter les $N\ Pr(M/NF)$ cas qui seraient survenus de toute façon en l'absence du facteur F. On peut espérer éviter les NF autres. La proportion maximale de cas que l'on peut éviter est donc :

$$\frac{Pr(M) - Pr(M/NF)}{Pr(M)}$$

Soit f la proportion de sujets exposés. En écrivant que : $Pr(M) = f\ \mathrm{Pr}(M/F) + (1\text{-}f)\ Pr(M/NF)$ on trouve que le risque attribuable vaut :

$$\frac{f(\mathrm{RR} - 1)}{f(\mathrm{RR} - 1) + 1}$$

Résultat :

Dans une population où une proportion f des sujets est exposée à un facteur F augmentant le risque de maladie, cette augmentation étant caractérisée par le risque relatif RR = $Pr(M/F)/Pr(M/NF)$, le risque attribuable au facteur, c'est à dire la proportion maximale de cas qui peut être évitée vaut

$$\frac{f(\mathrm{RR} - 1)}{f(\mathrm{RR} - 1) + 1}.$$

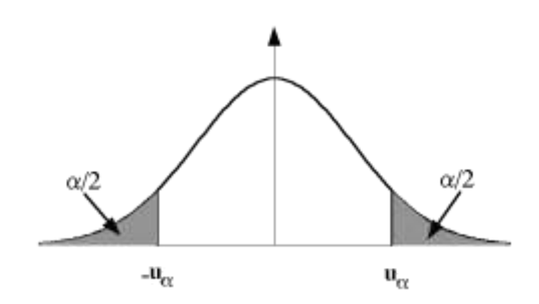
La proportion calculée grâce à cette formule est « maximale » : elle n'est atteinte que si le facteur F a un rôle causal dans le déclenchement de la maladie. Par exemple, l'exposition au tabac est un facteur causal du cancer des bronches. L'alcoolisme ne l'est pas. Pourtant, le risque relatif RR de cancer des bronches chez les alcooliques est supérieur à 1 parce que les alcooliques sont plus souvent fumeurs que les non alcooliques. L'alcoolisme est appelé facteur de confusion. Bien entendu, en supprimant l'alcool, on ne supprimerait pas le cancer des bronches !

Résumé du chapitre

1. L’essai contrôlé randomisé permet de mesurer de l’effet causal d’une intervention de santé, un traitement par exemple.
2. La randomisation qui consiste à tirer au sort l’attribution de l’intervention, permet d’assurer que les individus constituant l’échantillon sont comparables en tout (homogènes) sauf pour ce qui concerne le caractère contrôlé.
3. Dans un essai randomisé, le critère de jugement est la variable qui sera comparée entre les groupes pour juger de l’efficacité de l’intervention. On distingue critères de jugements objectifs (ex : décès) et subjectifs (ex : douleurs), ces derniers pouvant être facilement influencés par d’autres effets que les effets propres de l’intervention.
4. L’effet thérapeutique dans un essai est la somme de l’effet pharmacologique propre et de l’effet placebo.
5. La mise en aveugle qui signifie que ni le patient, ni le médecin qui le suit, ni l’évaluateur du critère ne savent dans quel groupe est randomisé le patient, est utilisée pour limiter les biais.
6. L’analyse en intention-de-traiter signifie que l’on compare le critère de jugement entre les groupes tels qu’ils ont été constitués par la randomisation. Elle implique que tous les patients randomisés sont conservés dans l’analyse.
7. Dans une étude d’observation, il n’est pas possible de conclure causalement, juste de mettre en évidence des associations entre expositions (par exemple fumer) et événement de santé (par exemple un cancer).
8. Les études d’observations visent à identifier les facteurs associés à des événements de santé ; il s’agit souvent de risques.
9. On distingue les études de cohortes, où les sujets sont répartis en groupes en fonction de leur exposition (ex : fumeur/non fumeur) ; les études cas-témoins, où les sujets sont répartis en groupes en fonction de la réalisation ou non de l’événement de santé (ex cancer/ pas cancer) ; les études transversales, où expositions et événements sont mesurés simultanément.
10. Une étude est dite prospective lorsque l’exposition est mesurée avant la survenue de l’événement étudié. Une étude est dite rétrospective lorsque la mesure de l’exposition survient après la survenue de l’événement.
11. Le risque relatif et le rapport des cotes mesurent la force de l’association entre l’exposition et l’événement de santé étudié. Ils valent 1 en cas d’absence d’association.

Annexe - Tables statistiques

A.1 - TABLE DE LA VARIABLE NORMALE REDUITE Z



α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	∞	2,576	2,326	2,170	2,054	1,960	1,881	1,812	1,751	1,695
0,10	1,645	1,598	1,555	1,514	1,476	1,440	1,405	1,372	1,341	1,311
0,20	1,282	1,254	1,227	1,200	1,175	1,150	1,126	1,103	1,080	1,058
0,30	1,036	1,015	0,994	0,974	0,954	0,935	0,915	0,896	0,878	0,860
0,40	0,842	0,824	0,806	0,789	0,772	0,755	0,739	0,722	0,706	0,690
0,50	0,674	0,659	0,643	0,628	0,613	0,598	0,583	0,568	0,553	0,539
0,60	0,524	0,510	0,496	0,482	0,468	0,454	0,440	0,426	0,412	0,399
0,70	0,385	0,372	0,358	0,345	0,332	0,319	0,305	0,292	0,279	0,266
0,80	0,253	0,240	0,228	0,215	0,202	0,189	0,176	0,164	0,151	0,138
0,90	0,126	0,113	0,100	0,088	0,075	0,063	0,050	0,038	0,025	0,013

La probabilité α s'obtient par addition des nombres inscrits en marge

Exemple : pour u_α = 0,994, la probabilité est α = 0,30 + 0,02 = 0,32

TABLE POUR LES PETITES VALEURS DE LA PROBABILITÉ

α	0,001	0,000 1	0,000 01	0,000 001	0,000 000 1	0,000 000 01	0,000 000 001
u_α	3,29053	3,89059	4,41717	4,89164	5,32672	5,73073	6,10941

A.2 - TABLE DU TEST DE WILCOXON

n	α		
	0,05	**0,02**	**0,01**
6	2,118		
7	1,961	2,299	
8	2,044	2,324	2,464
9	2,026	2,263	2,381
10	1,947	2,253	2,456
11	2,009	2,276	2,454
12	2,008	2,322	2,479
13	1,964	2,313	2,523
14	1,952	2,329	2,517
15	1,965	2,306	2,533

Indique, pour $n \leq 15$ les valeurs de W_α pour α = 0,05, 0,02 et 0,01.

(d'après Fisher et Yates, Statistical tables for biological, agricultural, and medical research (Oliver and Boyd, Edinburgh) avec l'aimable autorisation des auteurs et des éditeurs)

A.3 - TABLE DU TEST DE MANN-WHITNEY-WILCOXON

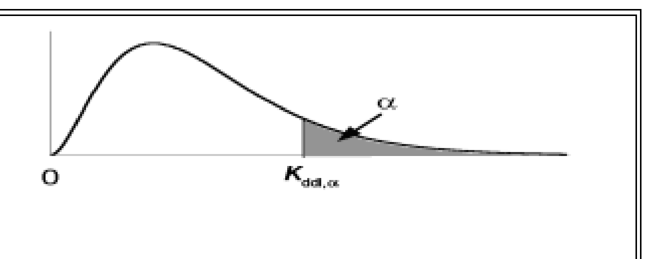
n_B	α	n_A 3	4	5	6	7	8	9	10
4	**0,05**	2,333	1,905						
	0,01	*2,687*	*2,483*						
5	**0,05**	2,117	2,107	2,110					
	0,01	*2,415*	*2,596*	*2,528*					
6	**0,05**	1,962	2,047	2,118	2,018				
	0,01	*2,479*	*2,473*	*2,483*	*2,498*				
7	**0,05**	2,074	2,003	1,965	2,086	2,057			
	0,01	*2,530*	*2,570*	*2,615*	*2,514*	*2,568*			
8	**0,05**	1,960	1,970	1,991	2,014	2,037	1,953		
	0,01	*2,572*	*2,480*	*2,576*	*2,530*	2,500	*2,584*		
9	**0,05**	2,052	2,099	2,013	1,956	2,022	1,982	2,040	
	0,01	*2,422*	*2,561*	*2,680*	*2,546*	*2,551*	*2,560*	*2,570*	
10	**0,05**	1,961	2,065	2,033	2,017	2,010	2,008	2,009	2,011
	0,01	2,366	2,489	2,523	2,560	2,498	2,541	2,580	2,540

Indique, pour $n_A \leq 10$ et $n_B \leq 10$, $n_A \leq n_B$, les valeurs de M_α, pour α=0,05 et α=0,01.

Exemple : n_A=5, n_B=8 : $M_{0,05}$=1,991

A.4 - TABLE DE χ^2

La table donne la probabilité α pour que χ^2 égale ou dépasse une valeur donnée, en fonction du nombre de degrés de liberté (d. d. l.)
Quand le nombre de degrés de liberté est élevé, $\sqrt{2\chi^2}$ est à peu près distribué normalement autour de $\sqrt{2(\text{d.d.l.})-1}$ avec une variance égale à 1

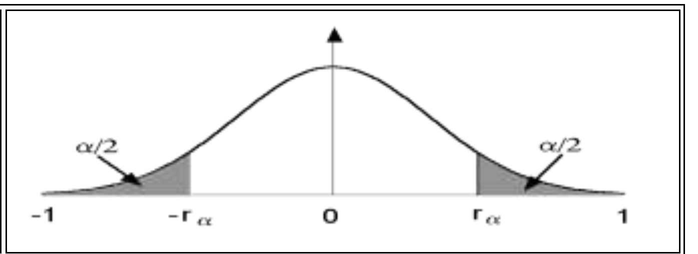


ddl / α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,0158	0,455	1,074	1,642	2,706	3,841	5,412	6,635	10,827
2	0,211	1,386	2,408	3,219	4,605	5,991	7,824	9,210	13,815
3	0,584	2,366	3,665	4,642	6,251	7,815	9,837	11,345	16,266
4	1,064	3,357	4,878	5,989	7,779	9,488	11,668	13,277	18,467
5	1,610	4,351	6,064	7,289	9,236	11,070	13,388	15,086	20,515
6	2,204	5,348	7,231	8,558	10,645	12,592	15,033	16,812	22,457
7	2,833	6,346	8,383	9,803	12,017	14,067	16,622	18,475	24,322
8	3,490	7,344	9,524	11,030	13,362	15,507	18,168	20,090	26,125
9	4,168	8,343	10,656	12,242	14,684	16,919	19,679	21,666	27,877
10	4,865	9,342	11,781	13,442	15,987	18,307	21,161	23,209	29,588
11	5,578	10,341	12,899	14,631	17,275	19,675	22,618	24,725	31,264
12	6,304	11,340	14,011	15,812	18,549	21,026	24,054	26,217	32,909
13	7,042	12,340	15,119	16,985	19,812	22,362	25,472	27,688	34,528
14	7,790	13,339	16,222	18,151	21,064	23,685	26,873	29,141	36,123
15	8,547	14,339	17,322	19,311	22,307	24,996	28,259	30,578	37,697
16	9,312	15,338	18,418	20,465	23,542	26,296	29,633	32,000	39,252
17	10,085	16,338	19,511	21,615	24,769	27,587	30,995	33,409	40,790
18	10,865	17,338	20,601	22,760	25,989	28,869	32,346	34,805	42,312
19	11,651	18,338	21,689	23,900	27,204	30,144	33,687	36,191	43,820
20	12,443	19,337	22,775	25,038	28,412	31,410	35,020	37,566	45,315
21	13,240	20,337	23,858	26,171	29,615	32,671	36,343	38,932	46,797
22	14,041	21,337	24,939	27,301	30,813	33,924	37,659	40,289	48,268
23	14,848	22,337	26,018	28,429	32,007	35,172	38,968	41,638	49,728
24	15,659	23,337	27,096	29,553	33,196	36,415	40,270	42,980	51,179
25	16,473	24,337	28,172	30,675	34,382	37,652	41,566	44,314	52,620
26	17,292	25,336	29,246	31,795	35,563	38,885	42,856	45,642	54,052
27	18,114	26,336	30,319	32,912	36,741	40,113	44,140	46,963	55,476
28	18,939	27,336	31,391	34,027	37,916	41,337	45,419	48,278	56,893
29	19,768	28,336	32,461	35,139	39,087	42,557	46,693	49,588	58,302
30	20,599	29,336	33,530	36,250	40,256	43,773	47,962	50,892	59,703

Exemple : avec d. d. l. = 3, pour $K_{3;\alpha}$ = 0,584 la probabilité est α = 0,90
(d'après Fisher et Yates, Statistical tables for biological, agricultural, and medical research (Oliver and Boyd, Edinburgh) avec l'aimable autorisation des auteurs et des éditeurs)

A.5 - TABLE DU COEFFICIENT DE CORRELATION

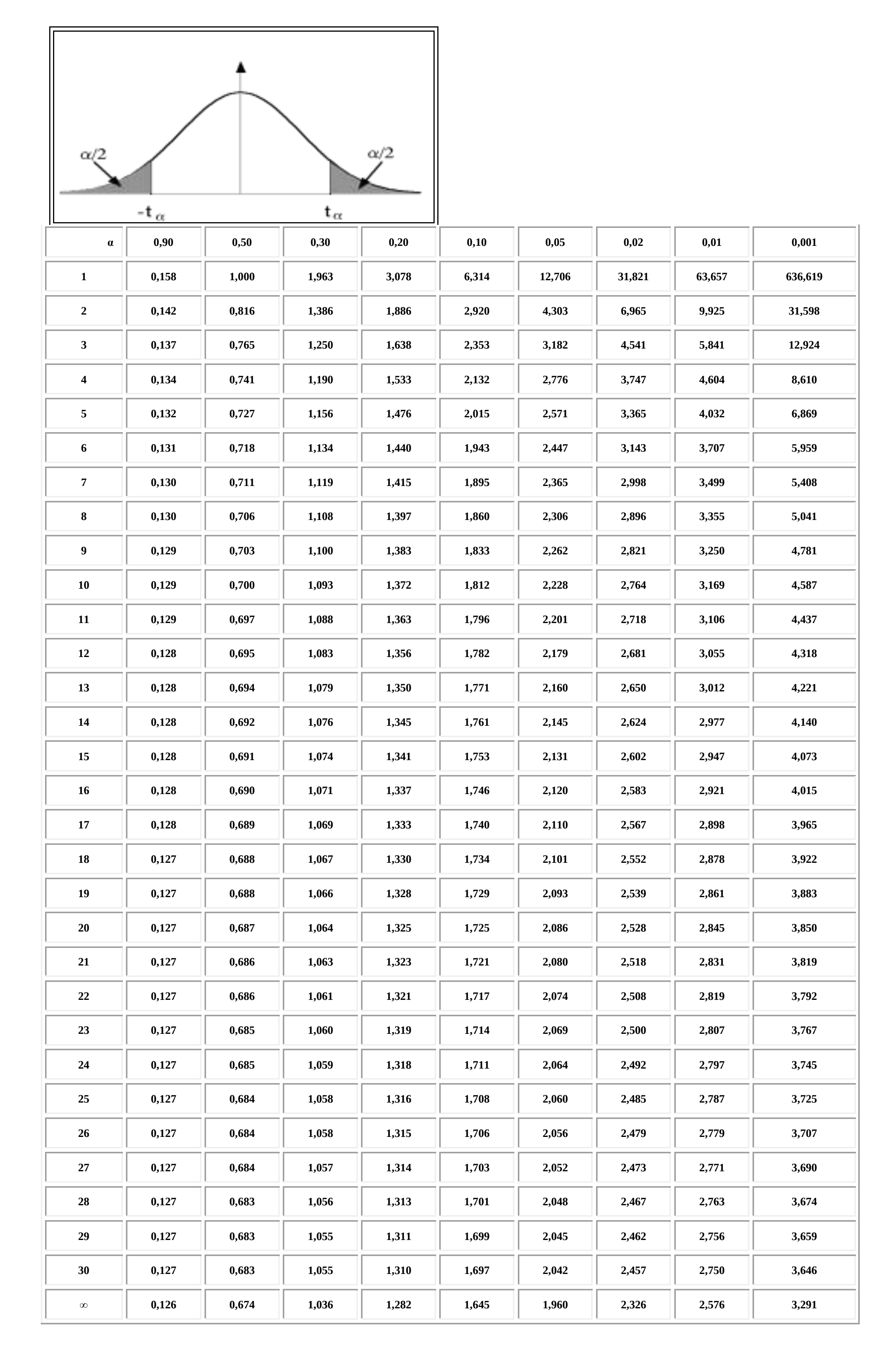
La table indique la probabilité α pour que le coefficient de corrélation égale ou dépasse, en valeur absolue, une valeur donnée r_α, c'est-à-dire la probabilité extérieure à l'intervalle ($-r_\alpha$, $+r_\alpha$), en fonction du nombre de degrés de liberté (d. d. l.)



ddl \ α	0,10	0,05	0,02	0,01
1	0,9877	0,9969	0,9995	0,9999
2	0,9000	0,9500	0,9800	0,9900
3	0,8054	0,8783	0,9343	0,9587
4	0,7293	0,8114	0,8822	0,9172
5	0,6694	0,7545	0,8329	0,8745
6	0,6215	0,7067	0,7887	0,8343
7	0,5822	0,6664	0,7498	0,7977
8	0,5494	0,6319	0,7155	0,7646
9	0,5214	0,6021	0,6851	0,7348
10	0,4973	0,5760	0,6581	0,7079
11	0,4762	0,5529	0,6339	0,6835
12	0,4575	0,5324	0,6120	0,6614
13	0,4409	0,5139	0,5923	0,6411
14	0,4259	0,4973	0,5742	0,6226
15	0,4124	0,4821	0,5577	0,6055
16	0,4000	0,4683	0,5425	0,5897
17	0,3887	0,4555	0,5285	0,5751
18	0,3783	0,4438	0,5155	0,5614
19	0,3687	0,4329	0,5034	0,5487
20	0,3598	0,4227	0,4921	0,5368
25	0,3233	0,3809	0,4451	0,4869
30	0,2960	0,3494	0,4093	0,4487
35	0,2746	0,3246	0,3810	0,4182
40	0,2573	0,3044	0,3578	0,3932
45	0,2428	0,2875	0,3384	0,3721
50	0,2306	0,2732	0,3218	0,3541
60	0,2108	0,2500	0,2948	0,3248
70	0,1954	0,2319	0,2737	0,3017
80	0,1829	0,2172	0,2565	0,2830
90	0,1726	0,2050	0,2422	0,2673
100	0,1638	0,1946	0,2301	0,2540

Exemple : avec d. d. l. = 30, pour r_α = 0,3494 la probabilité est α = 0,05(d'après Fisher et Yates, Statistical tables for biological, agricultural, and medical research (Oliver and Boyd, Edinburgh) avec l'aimable autorisation des auteurs et des éditeurs)

A.6 - TABLE DU t DE STUDENT



α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	1,000	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,816	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,765	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,134	0,741	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,727	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,718	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,706	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,703	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,700	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,697	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,695	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,694	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,692	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,691	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,690	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,689	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,688	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,688	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,687	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,686	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,686	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,685	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,685	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,684	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,684	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,684	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,683	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,683	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,683	1,055	1,310	1,697	2,042	2,457	2,750	3,646
∞	0,126	0,674	1,036	1,282	1,645	1,960	2,326	2,576	3,291

Exemple : avec d. d. l. = 10, pour *t* = 2,228, la probabilité est α = 0,05 (d'après Fisher et Yates, Statistical tables for biological, agricultural, and medical research (Oliver and Boyd, Edinburgh) avec l'aimable autorisation des auteurs et des éditeurs)